

THE CENTURY PSYCHOLOGY SERIES

RICHARD M. ELLIOTT, KENNETH MACCORQUODALE,  
AND GARDNER LINDZEY, *Editors*

PSYCHOLOGICAL  
RESEARCH

# *Psychological Research*

---

BENTON J. UNDERWOOD

NORTHWESTERN  
UNIVERSITY



*New York*

APPLETON-CENTURY-CROFTS, Inc.

Copyright, © 1957 by  
APPLETON-CENTURY-CROFTS, Inc.

*All rights reserved. This book, or parts thereof, must not be reproduced in any form without permission of the publishers.*

LIBRARY OF CONGRESS CARD NUMBER: 57-5112

PRINTED IN THE UNITED STATES OF AMERICA

E-89098

## PREFACE

THIS book deals with certain problems met in using the methods of science to study behavior. It has only one purpose, namely, to aid in training better scientists among those who make psychology their subject matter. I think there is evidence that we need to train better scientists in psychology; that this book will aid in attaining this objective rests only on a faith since I have no evidence on the matter. A strong need to train more competent scientists was felt by the staff at Northwestern University a number of years ago and among the steps taken to improve our program was the introduction of a course which dealt exclusively with research problems. This may or may not have been a wise step (introducing new courses cannot be the cure of all our deficiencies) and I mention it only to give a specific origin for the book. I was given primary responsibility for this course (listed in the catalogue as *Scientific Method in Psychology*) which is required of all graduate students in their first year. The present book consists of my lectures as they exist at the end of five years during which the course has been offered.

In writing the book (through teaching the course) I felt an obligation to reflect current research practices as I saw them, with emphasis largely on experimental research. As will be noted, I include under research practices far more than the design of an experiment and collection of data. Indeed, I have included topics which are distinctly controversial, and I have introduced issues which I think have been given far less attention than they deserve. The result is that in a certain sense the book becomes a philosophy of science. My philosophy of science, being as any philosophy is, a personalized affair, may not have allowed me to set down a true picture of the contemporary research scene. But, even if I were so unbiased that I could accurately reflect this scene, there are so many matters which are controversial

and on which I found it necessary to take a stand, that I fully expect to be disagreed with at several points. If I did not believe that the training of research workers would benefit from further discussion of these controversial matters, I would never have submitted these materials to the inspection of others.

To say that this work is a philosophy of science can only be said with an apology to those who are by profession philosophers of science, for this book is at best a pragmatic philosophy and may seem naïve to them. It is a pragmatic philosophy because it is concerned with issues and problems more closely related to the actual doing of research than are the issues and problems handled by the philosophers of science. Only at a few points do the problems clearly overlap those commonly found in the writings of the philosophers of science.

My debts are many. For over 10 years I have been privileged to be at a university which encourages and facilitates teaching and research in line with the finest traditions of our great educational institutions. I have also had the good fortune to be a member of a small department dedicated vigorously to research and teaching. Only a few of the present chapters have been read critically by my colleagues but I suspect every topic in the book has been discussed with me at one time or another by at least one associate. I mention this because while I recognize a real debt to my colleagues as a result of these discussions, it may be greater than I realize. The remarks which I boldly set forth as my own may have actually been germinated by one of my colleagues but the passage of time has obscured the source. Yet, it may be a blessing, for I know that my position on some matters is not popular and to attempt to give credit where the source is questionable might result in injustice.

Professors R. M. Elliott and Kenneth MacCorquodale have critically read the entire manuscript. They, too, have disagreed with my position on some issues but have left the final decisions to me. I owe both much for smoothing and tempering my prose.

Students who have listened to my lectures or read some of the materials have pointed out ambiguities and inconsistencies which I have tried to correct. Many of the illustrations in Chapters 3, 4, and 5 are taken from student reports. Mrs. Irene Nolte has typed the manuscript and has eliminated inconsistencies in the format.

Finally, I wish to thank the following publishing firms for allow-

ing me to quote material: University of Chicago Press; The Dryden Press; Appleton-Century-Crofts, Inc.; The Journal Press; Cambridge University Press; John Wiley & Sons; American Psychological Association; *American Journal of Psychology*; *American Journal of Physics*; *American Scientist*; *Science*.

# CONTENTS

	<i>Page</i>
PREFACE	v
<i>Chapter</i>	
1. INTRODUCTION	1
2. ANALYSIS OF THE RESEARCH SITUATION	17
3. OPERATIONAL DEFINITIONS	50
4. RESEARCH DESIGN: I	85
5. RESEARCH DESIGN: II	128
6. AN OVERVIEW OF EXPLANATION IN PSYCHOLOGY	174
7. SOME CHARACTERISTICS OF CONCEPTS	195
8. THE NATURE OF SOME EXPLANATORY ATTEMPTS	234
9. POTPOURRI	271
INDEX	293

## *Introduction*

### SOME GENERAL COMMENTS ON SCIENCE AND PSYCHOLOGICAL RESEARCH

The purpose of the methods of science is to achieve a description and understanding of nature (the universe). By description I mean the definition, cataloguing, or classification of events, objects, and phenomena which define nature, and the statement of empirical relationships associated with these events, objects, and phenomena. By understanding I mean the reduction to the smallest possible number of general laws which would account for the various specific facts. The descriptive part of science is concerned with research *per se*; what I have called understanding is usually achieved through theory.

This particular book is concerned with the scientific method as a means of studying behavior, particularly by those who call themselves psychologists. I will try to reflect faithfully various research practices in psychology; but, in spite of the manifest enthusiasm which I have for my profession, I find a great deal to criticize in these research practices. When I am critical it is in the interests of betterment of psychological research, not because of any overpowering urge to censure. For certainly, it seems to me, we need to maintain a continuous review or inspection of the attempts to apply scientific method to the study of behavior. Some of these attempts make science look ludicrous and they must be evaluated for what they are. Probably there is no other area of human endeavor which so badly needs a thoroughgoing application of the scientific method as does psychology, for probably in no other area are there so many misconceptions, so many half-truths, and so many abortive attempts to understand behavior.



The social forces following World War II markedly increased the number of psychologists in positions where (among other things) they are attempting to minister to the mental ills of mankind. Such ministrations are severely and perhaps critically hampered by an almost complete lack of relevant behavioral principles eventuating from research. To some this is a frightening situation. Whether the society of psychologists should have allowed themselves to be drawn into this situation is a complex and controversial matter; it is not my intent to debate the issues. But, the appalling schism between the facts of psychology and what many practitioners are trying to do is an issue, for the breach can only be reduced by more and better research. If hundreds of psychologists choose to work in industry, clinics, guidance centers, and so on, and if the profession is to survive as a respected one, there seems to be no answer except sound research, whether this research is done in the applied setting or in the universities.

We at the universities where graduate work is done are almost entirely responsible for the training of research people. We are largely responsible not only for the quality of research work but also for whether or not it is done. We cannot escape the responsibility we have of inculcating the highest standards of research in our students as well as training more students for research careers. We must not only institute these standards but we must also continuously police ourselves against any lowering of them. Standards of research are not static; the highest type of research standards at any given time almost inevitably leads to a subsequent raising of the standards; good research breeds better research. Society must not only be protected against the practitioner who operates without the leavening influence of principles derived from research but likewise must be protected against the shoddy effects of ill-conceived and grossly misinterpreted research.

It is apparent that universities vary considerably in standards of research for their students, and the diverse standards are disseminated in turn by these students to their students. I do not think there is a middle ground on these matters; psychological research is an honorable profession and training for such a profession must be at the highest level we know. An examination of current research reported

in our journals shows that the essential aspects of the scientific method are still largely foreign to some psychologists. The critical aspects of this book, therefore, stem from a faith that there must be ways by which general standards of research practices can be elevated, at least to some degree.

Now, you may ask: "Why this fixation or fetish on the application of scientific method to psychological problems?" The answer is that no one has conceived of a better way for demonstrating and understanding the lawfulness of nature. I, therefore, believe that we should promote with all our vigor the appropriate use of these powerful tools of understanding. I would not, of course, deny the right of novelists, poets, artists, or indeed, metaphysicians, to record their interpretation of human behavior. On the contrary, I would defend such a right so long as it is accurately described for what it is and the interpretation clearly distinguished from those based on scientific method. At least at the present time, the word "science" seems to have a certain prestige value, and we find the most curious activities masquerading as scientific endeavors. The record should be kept clear.

Some other preliminary remarks need to be made to set the tone of this argument. I will not engage in quarrels about the philosophical bases of science, about its social implications, nor its evils and virtues. In some instances I will make some assertions about these matters if I think they clarify subsequent material. I shall make no attempt to defend science and scientists against certain criticisms; if this need be done, it has been done (e.g., 2, 5, 7). My basic premise is that scientific research in psychology (as well as other disciplines) is a vital part of man's ever-extending endeavor to comprehend the universe. I wish merely to discuss critically some of the problems of research in psychology as I see them.

### THE ASSUMPTIONS OF SCIENCE

Probably not many scientists are able to formulate adequately the assumptions which logically underlie their labors. Moreover, it is likely that the average scientist has not done much thinking about these assumptions, for he can do perfectly good work without it. However, since these issues sometimes plague the curious student of

science, I want to discuss what I consider the two basic assumptions of science.

*Determinism.* One of the assumptions of a scientist is that there is lawfulness in the events of nature as opposed to capricious, chaotic, or spontaneous occurrences. Every natural event (phenomenon) is assumed to have a cause, and if that causal situation could be exactly reinstituted, the event would be duplicated. In the strictly physical world, determinism would probably be accepted by all, scientist and layman alike. Apples do not drop up one day and down the next; gasoline engines do not run without fuel; rain does not fall unless there are clouds in the vicinity. In short, there is a predictability (reliability) of the events in the physical world, and few would disagree that appropriate search would be likely to find the particular conditions under which the events occur.

Now certainly the awareness of orderliness and lawfulness in nature is not a product of modern science. Early man noted regularities such as the changes of the seasons and the growth of animals and plants under certain conditions and not under others. Science has merely allowed us to pyramid these regularities systematically and go back of apparent causes to more basic causes or correlates (antecedent events or causes); in so doing it has allowed us to bring many phenomena under a single causal principle. It has, in effect, ordered the orderliness of nature as distinct from casual observations and the unrelated interpretations of common sense. Furthermore, science discovers orderliness about phenomena which are not readily apparent to the human senses.

So, in general, the principle of determinism is an underlying assumption of the scientist, and for the apparent physical world is widely accepted. Even here, however, the acceptance is not universal, especially where topics such as the origin of life, or the doctrine of evolution are concerned. That the scientist does not always realize his acceptance of the principle of determinism probably stems from the obviousness of it (to him). It is taken so thoroughly for granted that verbalization of it would appear redundant. He applies his methods of science and time after time finds the orderliness of nature of which we have spoken. Indeed, even if he found chaos in a given area of nature, it is quite likely that he would not imme-

diately, or indeed soon, question the assumption of determinism. Rather, he would look to his investigative procedures for the reasons behind the discovery of chaos instead of orderliness. Since science has found orderliness decade after decade and in subject matter after subject matter, the scientist is prone to believe that all events of nature, be they events characteristic of stones, oceans, anglerfish, ministers, corpuscles, or nerve tracts, have discoverable correlates. But, whether or not the scientist has thought about or verbalized this assumption is basically irrelevant to his work as a scientist; it is quite possible for him to carry on excellent scientific work without ever having heard about determinism. (See Benjamin, 1, for a more complete discussion of why the scientist may not pay much attention to this assumption.)

When we turn to the problem of determinism in human behavior, the principle is not so easily accepted by all. There are people, educated and uneducated, prominent and obscure, who do not hold to the doctrine of determinism in human behavior. Certain religions can accept it up to a certain point only to abandon it beyond that point for other explanatory principles. I will not argue these matters; the interested reader is referred to a paper by Grünbaum (4). It is sufficient to say that to reject determinism for a part or all of human behavior is in a sense to reject the application of scientific methods to the study of human behavior. Rejection of such a fundamental premise at this stage of development of psychology is decidedly premature, for application of scientific methods by psychologists has already revealed a pattern of orderliness in behavior and many cause-effect relationships commensurate with the age of psychology as a science. There is plenty of room for pessimism about how rapidly the application of scientific method to the study of behavior will reveal all cause-effect relationships which are necessary for a fairly complete understanding of human behavior, i.e., to reach a stage that is roughly equivalent to the knowledge achieved by physicists. Even those of us who are the most ardent advocates of the use of the scientific method may have rare moments of despair when we realize how little progress will probably be made in our own lifetime toward a thorough understanding of the behavior of the human child or adult. But the motivation for discovery, whatever its

source, is remarkably resistant to extinction. Basically the research psychologist knows, however personally important his research may be to him, that his lifetime contribution to understanding behavior will be small. Yet in spite of this and in spite of the small extrinsic rewards likely to accrue from his efforts, he retains an unshakeable belief in the doctrine that all behavior, simple or complex, is determined by discoverable causes and will eventually yield to the methods of science. Determinism is a necessary assumption for the scientific enterprise.

*Finite causation.* A second general assumption made by the scientist is that every natural event or phenomenon has a discoverable and limited number of conditions or factors which are responsible for it. For, as Pap (10) indicates, science would be almost a hopeless undertaking if nature were so constituted that everything in it influenced everything else.

This assumption need not be dwelled on. The length of an astronomer's toenails doesn't influence the phases of the moon; the color of the secretary's hair doesn't affect the height to which the corn grows in an Iowa field, and a Pygmy tribe in New Guinea has little influence on the alcoholic consumption of a truck driver in Brooklyn.

*Specific assumptions.* There are many less general assumptions with which the scientist must deal in his day-to-day work that are probably more important from a pragmatic point of view than are the more general assumptions. Speaking now only of psychological research, there are many different methods by which a given problem may be attacked. Each of these methods involves certain assumptions, some common to all the methods, some unique. There are in addition the omnipresent assumptions dealing with sampling and statistical analyses. The research psychologist must weigh the seriousness of failing to meet certain assumptions; he must evaluate which method violates assumptions least, if at all; he must, in short, choose the method which has the greatest probability of meeting the assumptions of acceptable research procedure. I merely mention these matters at this point, and will deal with them no further here, for these are major problems which are reserved for later discussion in another section of the book.

## THE HISTORY OF SCIENCE AND THE HISTORY OF THE SCIENTIST

Scientific work is an unending series of analytical steps. If we ask about our understanding of the complete order of the universe, the series of analytical steps extends beyond our vision. Such a statement is real in the sense that it reflects the history of science, and, using this history as a base for projection into the future, we see no other picture except this series of analytical steps. But such a statement gives not the slightest hint of science as it appears in concrete form to the working scientist. There is something very cold, forbidding, and uninspiring about a description of scientific work as an infinite series of analytical steps. But, to the working scientist, to the man who takes his research seriously, it becomes at once the most stimulating, frustrating, exciting, discouraging occupation imaginable. The few analytical steps successfully taken by a scientist in his lifetime are interlarded with defeats, misconceptions, and bumbles. The analytical steps are the fruits which appear in the history of science. Only a detached historian, looking at science and not at the scientist, can describe science as calculated and cold. There is nothing chilling, ruthless, nor inexorable about the march of science to the scientist as he works. Science in practice is full of dead ends; plenty of its great discoveries occurred as if by accident; it has its share of blunderers as well as men of brilliance. Put the positive products of these men's endeavors together in the history of science and we have the long series of what might appear at times to be capriciously generated analytical steps.

The process of discovery has been as varied as the temperament of the scientists. Of course, individual research projects are as a rule unspectacular, within their small scope fairly routine and logically consistent; but precisely some of the most important contributions have initially depended on wrong conclusions drawn from erroneous hypotheses, misinterpretations of bad experiments, or chance discoveries. Sometimes a simple experiment yielded unexpected riches, whereas some most elaborately planned assaults missed the essential effect by a small margin. Great men at times had all the "significant facts" in their hands for an important finding, and yet drew trivial or wrong conclusions; others established correct schemes in the face of

apparently contradictory evidence. Even the work of the great heroes, viewed in retrospect, sometimes seems to jump from error to error until the right answer is reached as if with the instinctive certainty of a somnambulist; indeed, this gift must be one of the deepest sources of greatness (6, p. 90).

The student who decides to be a research scientist and who starts his work in that direction may have a misconception of the rewards of his occupation. Many of us thought as we first started doing research that next Monday, or next semester, or surely next year we would make revolutionary discoveries or have revolutionary insights. It seldom happens. We must be satisfied, indeed, be proud of little insights and little discoveries, for these are the lifeblood of a science.

The growth of science... does not depend only on a few great discoveries. It depends equally on that slow accretion of multitudinous small steps that furnish the bases for, and the necessary extensions of, those discoveries, and also on the correction of the even more numerous missteps continually being made (8, p. 25).

To have an original thought, to perceive a new relationship, to comprehend a complex relationship, or, to plot out a well-controlled experimental design—to do these things and to receive personal satisfaction for doing them make the life of a scientist. If a person cannot work long hours at research without feelings of martyrdom, science is not his occupation.

#### PURE AND APPLIED RESEARCH

In this section I wish further to set the temper of the chapters to follow by making certain assertions about problems which continually arise around the dichotomy of pure *versus* applied research. These problems seem always to have been present in varying degrees even among scientists living in a strictly academic atmosphere. But, with increasing sponsorship of research by government agencies, the problems have extended into the governmental administrative, hence political, domain. And of course, the issues have ever been present as a consequence of the hiatus in the understanding of the layman of the scientist's motives. In discussing this issue I shall first try to make

clear what I mean by the essential terms I will employ in this discussion.

I use the terms "pure" and "applied" merely to identify the ends of a crude continuum. This continuum is defined by the attitude of the research worker. At the applied end of the continuum, we have the research worker who asks himself questions about the manner in which the world (nature or social order) is functioning and does research concerned with these questions only if it appears that the product of his research will clearly and immediately modify the way in which the world is functioning. At the other extreme is the investigator who asks himself questions about the world, questions about why nature behaves as it does, and sets about to get the answers without any concern that they may be used to change the world. All this pure research worker wants to do is understand the world. In between these extremes, of course, are gradations. Without doubt there are many research workers who ask themselves research questions as a result of a basic curiosity about nature and then further ask what relevance the answers to such questions might have in changing the world. Whether they proceed with the research or not depends on the values they place on the two aspects of the problem. And of course, a man need not occupy a static position on the continuum; he may range as his interests and values change or, as during a war, when emergencies demand it.

Another facet of the pure-applied problem is that presented by the technologist. The technologist is one who applies the results of research; he uses the knowledge gained by research to change the world. The technologist may be a different person from the research worker, but it is also quite obvious that he may himself be engaged in research. That is, we may well have a research worker with a strong technological bias; he does the research and also applies the knowledge to effect some change in the world.

Now it should be clear that the intent of this book is to discuss research methods and procedures *per se*, whether this research is applied or pure. But, I find it necessary so frequently to defend freedom of inquiry in general, that I must make a number of other comments about this subject, since it bears directly on the applied-pure problem.

Freedom of inquiry is a reflected but integral part of our Consti-



tutional liberty in the social order; it has been singled out, fostered, and protected largely by the great universities of our country. I suppose that it may be difficult for the average layman to comprehend just how much freedom of inquiry means to a scientist. And I think also it is hard for the scientist to express to the layman why it is such an essential component of the research atmosphere. When the scientist may pursue his work, wherever it may lead him (providing no harm befalls others during the pursuit), without having to answer the question, "what good is this?", that is what I mean by freedom of inquiry. Such a situation is still maintained at most of our universities. I have been at Northwestern University for 10 years; never once have I heard of research being questioned by a dean, or other administrative officer, or a colleague, because the research worker had no answer to the question "what good is this?"; indeed, the question is never asked. The research might be questioned on a number of grounds, such as methodological adequacy; but never does the man have to defend his work against the charge that it has no immediately foreseeable application. It is this guardianship of freedom of inquiry which to many is the most magnificent tradition of our universities.

Seldom are direct, frontal attacks made on this freedom. The attacks, when they occur, are neither calculated nor obvious but nonetheless are to be reckoned with. Recently, for example, a book appeared dealing with methods of research in education, psychology, and sociology (3). Let me give some quotes from this book, which, in many respects has my admiration but which on this matter frightens me:

This criterion of importance, in choice of a problem, involves such matters as significance for the field involved, timeliness, and practical value in terms of application and implementation of results (p. 54).

Scientific work in education, psychology, and the social sciences in general has an especially urgent obligation to play a social role in rendering service to society and humanity (p. 54).

It is high time that the social responsibilities of scientists and of research workers be recognized and accepted (p. 54).

The research worker is not expected, as a general rule, to implement the results of his studies, however desirable this consummation may be. He is not even compelled to point out the practical application of his

findings, although this step seems essential, especially in the social sciences (p. 55).

It is apparent, in my opinion, that such teachings can have a constricting effect on the very freedom that allows them to be published. But, let us set the record clear. Our society has a perfect right to expect scientists to be good citizens, to be loyal to the institutions under whose protection they live, and, in general, to be like any other citizen on these matters. But this is quite a different issue from saying that social scientists should be spending their time in research which will solve the social and political problems of the world. Society has always had problems crying for solution; science has helped solve many such problems. But, it would be a curious contradiction and a patently dangerous situation if science must heed every demand made upon it. Not only would such a turn of events be contrary to freedom of inquiry but it might be extremely shortsighted. The urge to succor the momentary ills of society springs from a noble motive, but that it is the best means of eliminating the ills of generations yet unborn may be doubted. The scientist's fundamental responsibility to society is to utilize his freedom of inquiry to the utmost, pursuing his researches wherever they may lead him in his field of competence.

It seems to me, therefore, that the evaluation of the importance of a piece of research will depend on the philosophical and on-the-job contexts which prevail. Private industry might evaluate it in terms of a step up in production; a defense department administrator on the basis of whether it would be of value in the training of new recruits; a university professor in relation to the soundness of its approach and how much it advanced our understanding of nature. There is no universal answer to the question of whether or not a piece of research is important. There are purely administrative decisions concerning it, and these will differ depending upon the philosophical convictions and values of the person making the decisions.

Without doubt it is the extreme purist in research that laymen find most difficult to understand. He clearly is a fellow who is interested in knowledge for knowledge's sake. If someone wants to make something practical out of his work, i.e., if a technologist uses

his results, he has no objection but he isn't interested in doing it himself. The history of science shows that these purists have, unwittingly, made many fundamental contributions to our social order. One could list many names, such as Mendel and Faraday, who worked for the sheer pursuit of knowledge, or understanding of nature but whose discoveries were applied later by others in a practical manner. One can easily imagine situations in which pure research could make dramatic contributions to the welfare of the world. Suppose, for example, that in the biology department there is a fine old professor who has spent his entire life studying butterflies. In most parts of the world butterflies have little impact on the order of nature; they don't harass the farmer's crops; they don't seem to be needed to maintain balance in the insect world; at very best their contribution is an aesthetic one. But, supposing someone should discover that the polio virus is transmitted by butterflies. At this point the exact and detailed knowledge—the pure knowledge for its own sake—would become tremendously important socially, its application of the highest importance. Butterflies could be brought under control almost immediately because complete data were available on their reproductive habits, life history cycles, and so on.

Let us not, therefore, be hasty in evaluating the worth of any research; what may seem to be pure and socially worthless today may become highly significant tomorrow, next year, a hundred years from now, or perhaps never. But because the pure research worker may, by his experiments, discover fundamental facts of nature, we must maintain the institutions in our society which will encourage such research. It is the universities, given support for pure research by philanthropic foundations and certain agencies of the government, which will continue to be the chief protectors of this unrestricted form of inquiry.

I am sure there are those who extoll the virtues of pure research and who loudly proclaim their right to do it, not from the basic desire to seek knowledge *per se*, but as a socially acceptable cloak behind which to retreat from reality. But, even so, while we might condemn such intellectual snobbery and detachment, highly significant research might actually be accomplished by such a person.

I do not think we can deny the contributions made to our

social order by pure research. Nevertheless, we could certainly question whether or not this insistence on the right to do pure research is an efficient way for science to proceed. We don't know, for example, how many pure researches have been done which will never have any practical value and which are not particularly important for understanding of nature; undoubtedly there are thousands. It is quite possible that the application of scientific findings to the solution of world problems, whatever they are, might be much further developed if all scientists had a strong technological streak in them. I do not know how such questions can be answered with assurance. Many respected scientists (e.g., Oppenheimer, 9) firmly believe that we need both the extreme purist and the extreme technologist for most rapid progress; the two complement each other. Of one matter we may be sure—as long as the atmosphere of our society remains free as it is today and as long as bigoted men never find their way into offices of power over research, we will continue to have research workers at all points on the continuum I have described. The degree to which the research done by a man is pure or applied depends upon his attitude, and this attitude is a product of our culture. As long as our culture tolerates, no, not tolerates, but fosters this diversity of attitude we will continue to have the great range in the nature of research. I think it should be that way, not because I necessarily believe that it speeds up acquisition of scientific knowledge or that it may lead more rapidly to social progress. These questions have no answer for me. But, I would want this to continue because I think freedom is a fundamental premise of science as well as government.

### PREVIEW

In completing this introductory chapter I will give a fairly extended preview of the topics to be covered in the subsequent chapters.

I shall talk about phenomena and laws about those phenomena as being the basic data with which psychologists work. Many of these phenomena are given specific names, such as color shock, intelligence, extinction, brightness contrast, pitch, stereotypes, and so on. But whether named or not, I shall simply call all of them

phenomena. The basic purpose of research in psychology is to discover phenomena, variables which affect them, and the lawfulness of the effects. In the next chapter, therefore, we shall set up a simple conceptual system around the research situation so that various aspects of the situation can be discussed. We shall see that various components of the situation may in themselves provide special problems of research.

In our initial research attempts in an area, one of the first stages is to demonstrate a reliable phenomenon (or phenomena) and give it an operational definition. Since the operational definition forms the base of any scientific inquiry, we will need to spend considerable time on the matter of constructing or formulating these definitions. We shall further discuss their limitations and implications. My particular way of viewing an operational definition is that in simple form it becomes an experimental design. With this as background we will then move wholly into the area of research design.

The material to be presented on research designs will be both expository and critical. The procedure will not be that of presenting in detailed form the acceptable research designs for the many types of problems on which psychologists work. These are available in a number of sources. Rather, I will look at general types of research designs which are used, with but little attention to specific variations needed for particular research problems. I will not be concerned to any extent with the statistical problems of research design. The major effort will be a cataloguing of major research errors which are being made today, all of which will be extensively illustrated. It is apparent, then, that I will attempt to teach thinking about research design by first pointing out errors that are frequently found in the literature and then showing how they can be avoided. When this material is accompanied by a search by the student for such errors in published literature, I have found it to be highly instructive providing one does not allow the negative aspects to overshadow the basic fact that good sound research can be and often is done.

The material covered thus far in this preview of chapters to come is largely concerned with the descriptive aspects of our science, that is, with the problems associated with the discovery of phenomena and the working out of variables related to them. The critical

material on research designs is concerned with the establishment of reliable phenomena (and laws about them) and avoiding pseudo-phenomena. Now, as indicated earlier, the second general province of science is that of explanation, theory, or understanding in a more comprehensive way than that given by the immediate data. While there is much disagreement among philosophers of science, and scientists too, as to just what theory is, there is fairly general agreement on the objective of explaining by means of theory. This is to say that explanation is the reduction of all laws or relationships to as few as possible independent basic concepts and assumptions. It is, in effect, showing that specific or detailed empirical findings are special cases of (can be deduced from) more general laws. But, the methods and terminology used in carrying this out in psychology have produced a host of problems. Theoretical or explanatory attempts are extremely unstructured; there is an appalling lack of agreement on terminology; there are unresolved problems on when an empirical concept becomes a theoretical concept and vice-versa. There is the further problem of viewing the theorist as a human being, with scores of orienting biases. Nobody is as invulnerable as a theorist who does no research, and nothing is as impregnable as a theory which suggests no research. At the same time, nothing is so fatal to a theory as a well-ordered set of empirical relationships.

I must say quickly that I make no pretense of bringing order out of the chaos. My only hope is that we can develop a set of standards or a point of view by which we can approach this chaos with somewhat greater understanding. And of course, my concern will not be with assessing any particular theory and its relationship or adequacy to a particular subject matter. Rather, I shall use illustrative theoretical formulations to demonstrate the diversity of approach which is occurring in the attempts at explanation.

Finally, I have set aside a concluding chapter for presenting a number of ideas on research which have not been covered in other sections. Perhaps the word *ideas* is inappropriate; perhaps the term *biases* is more accurate. But, since the editors of this book have not seen fit to strike them out, they may at least serve the purpose of generating fruitful arguments.

## REFERENCES

1. BENJAMIN, A. C. Science and its presuppositions. *Scient. Mon.*, 1951, 73, 150-153.
2. FEIGL, H. The scientific outlook: Naturalism and humanism. In H. FEIGL & M. BRODBECK (Eds.) *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953.
3. GOOD, C. V., & SCATES, D. E. *Methods of research*. New York: Appleton-Century-Crofts, 1954.
4. GRÜNBAUM, A. Causality and the science of human behavior. *Amer. Scient.*, 1952, 40, 665-676.
5. HILDEBRAND, J. H. The social responsibility of scientists. *Amer. Scient.*, 1955, 43, 450-456.
6. HOLTON, G. On the duality and growth of physical science. *Amer. Scient.*, 1953, 41, 89-99.
7. MCCARTHY, H. E. Science and its critics. In P. P. WEINER (Ed.) *Readings in philosophy of science*. New York: Scribners, 1953.
8. MULLER, H. J. Science in bondage. *Science*, 1951, 113, 25-29.
9. OPPENHEIMER, R. Encouragement of science. *Science*, 1950, 111, 373-375.
10. PAP, A. Does science have metaphysical presuppositions? In H. FEIGL & M. BRODBECK (Eds.) *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953.

## *Analysis of the Research Situation*

Among the products of the scientist's work are certain conceptual structures around which he orders his thinking. They seem to help him keep his ideas and facts in a tidy state of affairs. One such simple conceptual structure which has had widespread repetition in psychology is the S-O-R conception. These letters may be thought of as standing for the gross components of the research situation in psychology, namely, stimuli, organisms, and responses. The stimuli may be distal or proximal; the responses immediate or developmental in nature. This gross analysis need not necessarily imply any strong theoretical or orienting bias; but, psychologists of different interests place varying emphases on the three components. Some psychologists look for functional relationships between stimuli and responses with comparative disregard for the relatively permanent capacities of the organism. Personality theorists, on the other hand, are largely concerned with identifying and characterizing these organismic capacities and traits. The physiological psychologist often takes the stimulus-response relationship as a starting point as he searches for the physiological mechanisms mediating the relationship. Regardless of our particular research interests or of our particular penchants for types of research tools, all three components of the research situation are important and will continue to be important in our quest for comprehensive laws of behavior. For the presentation here, where the emphasis is on experimental methodology, all three components must be discussed in detail. We will start with the response.

### RESPONSES

Human activities, behavior, or more simply, responses, constitute the universe of phenomena which psychologists describe and attempt to understand. I do not wish to put any restrictions on the magnitude



of the response with which we will deal; it may be thought of as being as "small" as an eyeblink or as "large" as social interaction. It has been common to speak of responses as being glandular and muscular activity. But, we do not often measure responses at this level; rather, we measure *products* of glandular and muscular activity. We note a verbal response; we count pencil marks; we measure latencies. The facts of the case are that the great bulk of the data of current psychological research reflects a gross but strict behavioral analysis, i.e., what the individual or group accomplishes or does, not what muscles are used nor what the chemical action of the cells is. This is not universal; Guthrie (11) has made strong pleas for actually describing muscular movements with comparative disregard for what these movements may accomplish. Many physiological psychologists are concerned with responses at the strict physiological level, such as nerve discharges or thyroid activity. Nothing is sacred about any level of description. Nevertheless, certain levels of response analysis seem to be more fruitful or useful at a given stage in the development of a science than do others. Or, it may not be that they are more useful but only that they are more in vogue, or perhaps, easier to accomplish. It should be clear that for me to say that the great bulk of psychological research is conducted at the gross behavioral level does not make it "right." I do not know how to judge the "rightness" or "wrongness" of such an issue; I am stating merely what seems to me to be a fair appraisal of how most research psychologists are behaving today.

#### SCALES OF MEASUREMENT

It would be presumptuous (and pointless) to attempt to catalogue the myriads of responses which form the raw data of the psychologist and define behavior as he studies it. One needs only leaf through a few representative journals to realize the ingenuity shown by psychologists in selecting behavior segments for study. I will not at this point, nor at any later point, engage in disputes concerning the *significance* of the multifarious responses studied by psychologists. That is, a criticism which has been levied at psychologists (sometimes by psychologists, e.g., 2) is that the responses they study are not *really* important; they are not the responses which represent

the behavior of an army general weighing the prospects before sending armies into battle; they are not the responses involved in a lynching or a revival meeting; they are not reflective of the behavior of a Senate floor leader trying to get a bill through the chamber. However, I *would* defend the proposition that research in psychology necessarily involves measurement, and that the rapidity with which research will embrace these so-called significant behaviors depends on our ability to break them down into relevant parts which can be measured.

There are a number of sources in which one can find detailed evaluation of scales of measurement extant in psychology (e.g., 6, 22), and no attempt will be made here to reproduce these discussions. Rather, I will simply indicate the diversity in the nature of scales in use and note a few points most pertinent for subsequent discourse.

The crudest level of response measurement used in psychology is that of identifying responses as belonging to one of two mutually exclusive categories. The basic data, therefore, consist of the frequency of responses in each category. Thus, we might make a tally of the number of students who did pass a particular course and the number who did not. Or, we might count the number of people who have visited a psychiatrist and the number who have not. This is measurement in its crudest sense, but nevertheless it is fundamental to all more precise forms of measurement and has itself been used many, many times in psychological investigations. Now, while it might be apparent that research in psychology is not concerned primarily with such counting, as an end activity, it may be worthwhile to make this explicit. I suppose that acquiring the knowledge that 85 per cent of the populace have never visited a psychiatrist while 15 per cent have has a certain significance in and of itself. But, a research psychologist would be interested in discovering in what other characteristics the two classes of individuals differ. That is, what are the correlates of visiting or not visiting a psychiatrist? Of course, in this illustration, one obvious difference would be expected, namely, mental illness. But, there may be a whole host of other factors which are related, such as *family background, financial status, age, and so on*. In short, defining a response measure (behavior) is

only an early step in research, and the two-category type of response measure is the crudest used.

Now note that assigning organisms (as a result of their responses) to one of two classes may involve *only* a single response; indeed, in the simplest case they are assigned on the basis of the presence or absence of the response. Members put in the same class may differ markedly on many other responses. For example, to take an extreme case, assume that for some esoteric reason a research worker in speech wanted to assign organisms to one of two classes depending upon whether they can or cannot be taught to speak. Thus, people and parrots would fall in the same class.

I have said that classifying responses into one of two categories is the crudest form of measurement. Once we go beyond two categories we begin to introduce a refinement in our response measures, for the members of a given class become more homogeneous. For example, we might classify all responses on a Rorschach test as form responses and not form responses. But, we might go further and classify them in terms of color, form, movement, or none of these. As our categories become greater and greater in number the responses in those categories become more and more homogeneous. If you like a name for this simple classifying-type measurement, the term *nominal* is commonly used (22).

Let us turn next to the other extreme. If a response measure is recorded along a *physical* or *ratio scale* we have the most advanced form of measurement. Thus, length, weight, time, and so on, are measured along such scales. If reaction time is measured, it is in terms of a ratio scale; kilograms of work per unit of time reflects the use of two ratio scales. Such scales, it will be realized, have a true zero and theoretically can be broken down into an infinite number of equal units.

In between these two extremes of precision of measurement we have several gradations of precision. Let us note first that in the nominal scale (yes or no, is or ain't, classification) we simply report the presence or absence of a response. Implicit in most of these classifications is the idea of magnitude of response. Suppose we classified adults as teachers or not teachers. In a certain sense, when we do this, we are saying that those we classify as teachers have a positive amount of "teacherness" and the other group has none. Our response

measure may make this idea of magnitude explicit and responses are rank ordered in terms of amount or magnitude of them. For example, we could rank-order 10 teachers as to their general teaching ability. This doesn't tell us anything about how much better Teacher A is than Teacher B, nor B than C, but the idea of magnitude or amount of teaching ability becomes explicit by such a measuring instrument. As we advance further up toward the physical scale, we may acquire instruments which will not only rank order but which will also give us information concerning the magnitude of the differences between ranks.

I will not pursue these matters further; as indicated earlier, excellent sources giving detailed differences among types of scales are available. There are, however, three additional points which I would make by way of setting up background for material to be covered later.

1. As seen above, there are certain responses which are studied by psychologists which can be described by physical or ratio scales. On the other hand, there are many responses for which no physical scales are appropriate. When a given response cannot be described by a physical scale, it must be measured by notations reflecting directly the discriminatory or perceptual responses of humans. If a characteristic of behavior can be described by perceptual responses in such a way that it is shown to vary systematically in amount, we have, I shall assert, demonstrated the existence of a *psychological dimension* and the instrument used to mirror this dimension is a *psychological scale*. The minimum requirement for establishing such a psychological dimension is two points; that is, Response A must be judged consistently to have a greater magnitude (or some other amount-like term) than Response B.

2. The greater the number of useful units in our scale (whether physical or psychological) the better (more precise) will be our response prediction. By useful I mean units which will consistently reflect differences in behavior. This is a problem of reliability to which I will return more fully at a later point.

3. Constructing psychological dimensions should not be thought of as merely being a useful technique for response measurement. In a general sense, the quantification of characteristics of objects or behavior via the human discriminatory response makes these char-

acteristics available for further research as manipulable stimuli conditions. On this point also I will speak at length in a later section of the chapter.

### RELIABILITY OF RESPONSE MEASUREMENT

Once we have identified the dependent variable, the response which we wish to investigate, and once we have determined the some level of quantification is feasible, the next step is ascertaining the reliability of the scale. Unfortunately, this is a matter which too many of us overlook and yet as far as the actual research is concerned, it is of the greatest importance. If our response measure is not reliable, no further investigative procedures should be undertaken. Science attempts to discover and understand reproducible phenomena; lack of reliability in our attempts at measurement precludes this reproducibility.

In the field of psychology I think those who construct tests have been far ahead of the strict experimentalists on this matter of reliability, the clinicians in general considerably behind. When a paper-and-pencil test is constructed, about the first thing the investigator does is determine its reliability, and the index is usually high for such materials. It may be that the investigator cannot find any other behavior which is correlated with the test behavior, but, by golly, he knows that whatever he is measuring he is measuring consistently. Other comments apropos to reliability may be incorporated in some illustrations.

Too often we take reliability for granted; I think we are likely to do this most readily when some equipment or mechanical instruments are involved. Instruments seem to have a halo of precision about them which tend to make us take their reliability for granted. But, in the current widespread use of electronic equipment, subject to continual breakdown, reliability must be checked repeatedly. Let me give you one illustration of the necessity of this.

During the later part of World War II, Air Force psychologists developed a gunsight which was to be used as a test for selecting men who would have high probabilities of becoming good gunners. This sight, an actual replica of the sight used on the B-29 at that time, gained its face validity because the subject actually aimed and

fired at a projected airplane as it flew across a curved screen. However, on the initial attempts to obtain normalizing data, the device was shown to be unreliable. An airman would come into the laboratory one day and make a good score and the next day a very poor score; day-to-day reliability coefficients ran about zero. Subsequent research revealed that the difficulty was not in the subjects (who might have been expected to evince variable motivation), nor in the device at any given session. Rather, the fault appeared to lie in the lack of constancy of calibration. The sight was constructed with very sensitive apparatus, and as a consequence of running it there were gradual shifts in the sensitivity of the scoring over the day's work. Hence, during one hour the device might score at a very sensitive level, whereas three hours later it would score at a very crude level. When constant checks were maintained on the calibration the reliability increased immediately to an acceptable level.

Among the more recently developed methods of recording responses is the "human yardstick method." The basic idea of the method has long been incorporated in the rating-scale technique. However, it has recently been applied in experimental studies of social and clinical behavior, the measurements being made at the time the subject is behaving. In the method, the judges are used to evaluate the behavior along certain specified dimensions. Thus, in a frustration experiment, the judges may be asked to rate amount of aggression, amount of motivation, and so on, at different points throughout an experimental session. Of course, certain experimental variations are introduced with the expectation that these conditions will produce differences in behavior which will in turn be reflected in the judges' ratings. In most instances where such research has been carried on, the investigators have faithfully calculated both intra-judge and interjudge reliability. Obviously, if the judges do not agree well on the amount of a response (such as aggression) manifested in the subject's behavior, the response measure has little usefulness. This in fact has sometimes been the case, although for other response measures reliability has been high. More details on such problems can be found in several sources (e.g., 14, 24).

I think I need only mention the high premium which must be placed on the reliability of scoring projective tests. If these tools are used as response measures for experimental purposes, the responses

acteristics available for further research as manipulable stimulus conditions. On this point also I will speak at length in a later section of the chapter.

#### RELIABILITY OF RESPONSE MEASUREMENT

Once we have identified the dependent variable, the response, which we wish to investigate, and once we have determined that some level of quantification is feasible, the next step is ascertaining the reliability of the scale. Unfortunately, this is a matter which too many of us overlook and yet as far as the actual research is concerned, it is of the greatest importance. If our response measure is not reliable, no further investigative procedures should be undertaken. Science attempts to discover and understand reproducible phenomena; lack of reliability in our attempts at measurement precludes this reproducibility.

In the field of psychology I think those who construct tests have been far ahead of the strict experimentalists on this matter of reliability, the clinicians in general considerably behind. When a paper-and-pencil test is constructed, about the first thing the investigator does is determine its reliability, and the index is usually high for such materials. It may be that the investigator cannot find any other behavior which is correlated with the test behavior, but, by golly, he knows that whatever he is measuring he is measuring consistently. Other comments apropos to reliability may be incorporated in some illustrations.

Too often we take reliability for granted; I think we are likely to do this most readily when some equipment or mechanical instruments are involved. Instruments seem to have a halo of precision about them which tend to make us take their reliability for granted. But, in the current widespread use of electronic equipment, subject to continual breakdown, reliability must be checked repeatedly. Let me give you one illustration of the necessity of this.

During the later part of World War II, Air Force psychologists developed a gunsight which was to be used as a test for selecting men who would have high probabilities of becoming good gunners. This sight, an actual replica of the sight used on the B-29 at that time, gained its face validity because the subject actually aimed and

reliable, we have only the first step in a research procedure, for certainly a reliable response measure is not very useful unless it can be shown to be related to something. But, in the above illustration, the low correlation between the two indices of tension cannot be attributed to the low reliability of the basic response measure (the tension index derived from the written protocols) although it might be attributed to the low reliability of psychiatric judgments.

As indicated earlier, considerable research effort goes into the dimensionalizing of characteristics of objects or symbols when these objects or symbols are to be used in subsequent investigations. There are many illustrations. An attitude scale may be constructed by having judges sort statements of opinion into piles along a defined continuum, the ends of which represent the extreme of the attitude involved, and the middle a neutral attitude. If a number of such items or statements can be shown to have high reliability, i.e., if the judges agree on the "degreeness" of attitude implied by the written statement, and if the dimension is well represented by statements having high reliability, the scale can then be used as a response-measuring instrument, say, in investigating conditions which might change the attitude. In verbal learning studies it is often necessary to dimensionalize certain characteristics of the material before proceeding with experimental work. Such characteristics as similarity, meaningfulness, familiarity, and affectivity have been dimensionalized by judges, and if reliability of the judgments is obtained the material can be used in subsequent research to discover the effect of the characteristics on various learning phenomena. While most of the scales have been constructed for work with human subjects, it is quite feasible to carry out operations whereby lower animals serve as "judges." For example, Harlow and Meyer (13) dimensionalized attractiveness of five different foods for monkeys by a paired-comparison technique. Knowing the value (to the monkeys) of each food, these foods can then be used in subsequent investigations to determine the effect of the values on certain learning behavior.

We need not pyramid the illustrations. In all cases, to repeat, the initial usefulness of the response measures depends upon the reliability of these measures.

*What is acceptable reliability?* It is by now quite apparent, I think, that in my opinion the measuring of response reliability is manda-



they evoke must be handled with the same attention to measurement problems as any other instrument. From these tools the raw data derived consist of written protocols. The investigator must then categorize what is believed to be relevant remarks or ideas in the protocols. If he is reliable within himself in this categorizing and if this is not an artifact (as it may well be), the response measure has usefulness. However, for research purposes, the use of several judges showing interjudge reliability is to be preferred.

The use of written documents to obtain response measures believed to be reflections of important aspects of behavior is somewhat on the increase. Since it is a somewhat unusual method of securing measurements of behavior, an illustration will be given. In 1947 Dollard and Mowrer (7) published a technique for deriving an index of tension from written documents. Essentially the technique consisted of analyzing a passage written by a patient by counting the number of clauses or phrases which to these investigators implied high tension or anxiety and the number which implied low tension or anxiety. A ratio between these two measures was used as an overall index of anxiety. In this particular article the authors were not concerned with what this tension measure might be related to; they were interested only in presenting the method and in showing that a reliable response measure could be derived. This they did. Inter-correlations among 10 independent scorers were quite high, much to the surprise of the scorers themselves who felt that the method was so subjective that little agreement would be evident. In another more recent study by Meadows *et al* (17) the reliability of such a response measure was again shown to be high, both when assessed in terms of different passages written by the same subject and in terms of the counting or evaluating by the investigators. Thus, the response measure was shown to be a consistent one for the same subject for different passages and also that different judges would derive about the same index for the same subject. These investigators also showed that the relationship between this measure of anxiety or tension and psychiatric judgment of anxiety made after interviews with the patients were correlated about zero. This discrepancy clearly produces a problem, a definitional one basically, but consideration of such problems will be delayed until later. At the present I simply wish to reiterate that even though a response measure is

ciation value of nonsense syllables. Even with a fairly limited number of responses for each subject the reliability was .79. One of my colleagues (8) built an apparatus a few years ago which required simultaneous activity of both arms in adjusting two levers. Although the mechanics of this apparatus were fairly complex, the odd-even trial correlation over 20 trials was .97. So, I would say, that while we must insist on reliability indices for new response measures, only rarely will they be low if the investigator has any "feel" at all for the factors in the situation which might introduce extreme intra-subject variability.

#### MULTIPLE RESPONSE MEASURES

There are four matters of widely varying importance regarding multiple response measures which need discussion. These four are: (a) the situation in which a research procedure yields two or more response measures which are highly correlated; (b) the situation which produces two or more response measures which are poorly correlated; (c) response correlation as an independent technique of research, and (d) response correlation and causality.

*Multiple response measures highly correlated.* There are many research situations in which the investigator records more than one response. For example, in learning studies it is quite common to record trials to learn, errors, and perhaps some other measure, such as latencies. Given this type of situation, we are concerned at this point only with the case where these response measures are highly correlated. Actually, little need be said about this. If two or more response measures are highly correlated, obviously we can use them to define a single phenomenon (some might prefer to say that we would infer a single process). In fact, if we have shown that we have two or more such measures, it becomes a rather redundant procedure to continue recording both in subsequent investigations. We could instead use any one of the measures with high confidence that we are measuring a single phenomenon. Which one we use becomes largely a matter of personal choice. If the recording of one response involves an expensive piece of apparatus and the other doesn't, our choice is made (unless we have a government research contract); in general we would choose the one which is most economical and

tory. It is, therefore, quite a legitimate question to ask when a response measure is reliable and when it is not. How large must a reliability coefficient be before the response measure can be accepted as a useful index of behavior? Unfortunately, a categorical answer to such a question is impossible. When a paper and pencil test is developed there are likely to be lifted eyebrows if the correlation coefficient is not at least .80 or more. But, we are all aware that the numerical value of the correlation is affected by several factors over which we have little control. For example, it is common knowledge that the use of a very homogeneous population will usually reduce the correlation as compared with a heterogeneous population. Split-half *versus* a test-retest technique is another matter affecting the correlation. Also, we know that if the response index has a very limited range imposed by the nature of the task the numerical reliabilities will be low. For example, in verbal learning studies, interday reliabilities of recall may run no higher than .50 and may be as low as .10. This would hardly seem to be satisfactory as a response measure even though these values differ significantly from zero. However, if one examines the situation it is discovered that the range of scores possible on such recall tests is so limited that individual differences cannot be fully reflected. That is, there may be 10 possible items to recall but because of the particular conditions of the experiment the total range recalled may vary from, say, 3 to 8. It is nearly impossible statistically to produce high coefficients of reliability from such data. The reliability coefficient of such response measures must be supplemented by the lawfulness of results which can be produced from experiment to experiment.

This is enough on response reliability. When a new response measure is used, or when an old one is markedly modified, we must make it common practice to derive an index of reliability. When the reliability is established, and only then, can systematic research be undertaken. Perhaps I have made too much of this issue. Actually, for the usual type of response measure, the reliability is likely to be high. And, I might say that there is something quite comforting in devising a new test or task and finding the reliability of the performance on this task to be very high. One of my students (19) recently worked out a technique to measure the associative capacity of subjects, a technique quite similar to that used in determining the asso-

measures have been used to infer strength of conditioning, such as amplitude, latency, frequency, and resistance to extinction. Various experimenters have used these measures interchangeably without first determining that they were highly correlated, hence, equally satisfactory to reflect the amount or strength of conditioning. And correlations between some of these response measures may be very low (3, 12).

What position are we in when we use two response measures to infer a single phenomenon when these response measures are poorly correlated? It is primarily a definitional or conceptual problem. It is a definitional or conceptual problem because if we have two poorly correlated response measures we actually must have two fairly independent phenomena and these should be so defined. Failure to do so may have at least two important complications. Not only can theoretical systems based on wanton interchange of such measures be quite misleading, but also I suspect that a number of so-called empirical contradictions in the literature may be a consequence of the fact that different response measures were used by different investigators and these measures were not highly correlated.

*Response correlation as a tool of research.* Response correlation may take at least three different forms as a complete tool of research. These forms are not independent, but I wish to mention them separately since they emphasize somewhat different rationales.

1. *Simple test validation.* Whether a test be a paper-and-pencil test, a performance test, or a projective test, the aim is that of predicting the behavior of the individual in situations other than the test situation. An investigator (for example) constructs a test which he believes will pick out potentially good supervisors from potentially poor supervisors. To get an index of validity he correlates test performance with subsequent supervisory performance. Or, an army psychologist, interested in predicting marksmanship performance, might correlate steadiness and marksmanship to see if the two are stemming in part, at least, from a common process or processes. As we all well know, predicting complex performance such as supervisory success or vocational success does not come easily. However, for our purposes the success or failure of such ventures is not particularly relevant. The germane point is the intent of the investigator in using response correlation. What he attempts to do is, by

practical to work with, reliabilities being equal. Nothing prevents us from continuing to record both measures, but little is gained by it.

I will give just one actual illustration of what I mean by high correlation between multiple response measures. Marquis (16) was interested in measuring reactions to frustration shown by newborn infants. The infants were observed for 10 days in the hospital immediately after birth. To attempt to produce frustration, the infant was allowed to have a bottle of milk for a short period of time, then the bottle was withdrawn, given again, withdrawn, etc. During the intervals between the short feeding periods, five different responses were recorded: (a) amount of mouth activity; (b) amount of general bodily activity; (c) frequency of crying; (d) latency of mouth activity; and (e) latency of general bodily activity. The lowest inter-correlation among these various measures was .80, even with a very small number of subjects. It seems apparent, that any one of these could be used as an index of frustration without fear that significant data were being lost by not recording the others.

*Multiple response measures poorly correlated.* In some research situations several response measures are recorded which are not highly correlated. Thus, responses to a Rorschach inkblot are categorized as movement responses, form responses, and so on. Presumably, these categories have low intercorrelations. This means that the investigator is simultaneously measuring different phenomena (or processes, if you prefer); a high frequency of movement response ostensibly means quite a different thing from a high frequency of form responses. In verbal learning experiments, rate of learning and frequency of overt errors have no relationship—the correlation is zero. Apparently, different mechanisms or processes are involved in the production of the two response measures. Actually, no immediate problem is evident if it is clearly shown that the response measures from a given situation have low intercorrelations and if the experimenter, therefore, concludes that he is dealing with different phenomena.

A problem with low intercorrelations among response measures may arise, however, if the response measures are obtained in different experiments and if an investigator uses them to infer the same process or define a single phenomenon. This problem has become a real one in the case of certain conditioning data. Several different

in rote learning. Both response measures were shown to be reliable, but no relationship was evident between amount of childhood punishment and error frequency. Thus goeth much research.

3. *Factor analysis.* The most refined and grandiose method of using response correlation as a basic instrument of research is that method known as factor analysis. Briefly, a group of subjects is given a large number of tests, say 20 or 30 which are presumed to sample all the various skills or capacities in a given domain, e.g., perceptual skills or motor skills or creative activity. Usually the tests are selected only following long periods of study of the various kinds of behavior believed to be involved in the domain selected for investigation. After the scores are obtained, correlations are calculated between each test and every other test; then, to speak facetiously, axes are rotated, matrices are matrixized, variances are variated, and vectors are vectorized. The purpose of this labor is to find out groups of tests which correlate highly with each other but not with other groups. Such a group of tests is therefore presumed to be measuring pretty much the same capacity or trait (factor) and will usually be given a name. As a consequence of such analyses it may be found that most of the variance can be accounted for by perhaps 4 or 6 essential factors. On subsequent research or in actual selection procedures only tests which are relatively purified for these factors may be used. We can see, therefore, that factor analysis results in an economy in that it identifies the skills which are important in a given area (domain) of behavior so that on subsequent research the testing becomes very limited.

There can be no doubt about the general usefulness of factor analysis as a descriptive tool. For the particular battery of tests used, the resulting factors define the capacities or skills involved. In a real sense it results in subject skills being defined by tests so that these skills have the status of a construct representing a broad area of capacity. If we then so wished on subsequent research we could manipulate these subject skills. To this matter I will return later.

The deficiencies of factor analysis (aside from any mathematical or statistical questions) lie not in the program of research outlined by eminent factor analysts, but in the failure thus far to carry out such a program. Thus it is asserted that once factors have been de-

abbreviated observations of behavior, predict behavior in a distinctly different situation. He may or may not be interested in the processes or skills as such which lie back of the performance; his interest may lie only in practical prediction.

2. *Personality correlates of nonclinical behavior.* I can best tell you what I mean to include under this rather inept title by giving you an illustration of the research which typifies what I am thinking about. It is represented in two theses done in our laboratory. An experimenter who has observed many subjects learn lists of words by rote cannot help but become curious about the personality variables or traits involved in such learning. It is a matter of fact that very little research has been done on this problem. By personality variables I am referring here to broad and rather loosely identified traits such as dominance or introversion. In rote learning, a phenomenon which attracted our attention as a possible personality indicator was number of overt errors. Some subjects make many errors, others very few, and there is no relationship between error frequency and rate of learning. It, therefore, seems quite conceivable that error rates reflect personality differences which are relatively independent of learning ability. More than that; one can generate specific hypotheses about errors and personality traits. For example, a subject who makes very few errors might be suspected of having a history in which he was severely punished (at home or school) for making errors. He might, therefore, have developed a generalized trait of caution against making responses unless he was fairly sure they would be correct. A person brought up in a loose disciplinary environment might, on the other hand, be relatively unconcerned about his errors.

To see if it were possible to find personality variables or traits related to error-making, Elkin (9) had 125 subjects learn by rote a rather difficult list of adjectives, and also gave them the Minnesota Multiphasic Personality test. The question was simple: will any of the various scales of the MMPI correlate with error frequency? None did.

In a second study by Singer (21) an attempt was made to test the specific hypothesis relating punishment in the subject's past history to error-making. By a questionnaire, Singer got information on the subject's history of punishment and also recorded error frequency

*Correlation and causality.* There is no agreement in science on what the appropriate use of the term *cause* really is. Some scientists even refuse the use of the term and prefer instead to speak merely of relationships. While I can appreciate the uneasiness attending the use of the word, it has a certain communication value and so I will not avoid it. For the time being I shall simply point out some difficulties which arise when the word is used in connection with interpreting the results of correlation studies. My initial point (one which has been made by many other writers) will be that inferring causality from simple correlations is an extremely dangerous pastime. Some rather extreme illustrations of this will be cited and then I will qualify the conclusion somewhat by looking at the way certain factor analysts conceptualize the problem.

If we find that Form L and Form M of the Stanford-Binet Intelligence Test correlate .95, I am sure that no one would conclude that the behavior observed on one form caused the behavior on the other. We might be willing to say that some hypothetical capacity or skill (intelligence) was measured about equally by both tests and that this capacity or skill is the immediate source (cause) of the observed correlation.

Suppose we notice that there is a high correlation between the number of people wearing raincoats and the amount of water in the storm sewers. We would not say that *because* people wore raincoats the amount of water in the sewers increased; nor would we say that the great amount of water in the sewers caused the people to wear raincoats. Obviously there is some other factor which is responsible both for the raincoats and the water in the sewers.

Some practitioners of factor analysis have found it easy and useful to think of their factors as causes. Thus, if a boy has high numerical ability as shown by test scores, and if he gets a high grade in a course in arithmetic, there is a tendency to think of the high grade as being caused by high number ability. Discussions by both Cattell (4) and Eysenck (10) make it clear that basically they would like to impute causal status to factor-analytic factors but cannot do so with confidence until they find independent conditions which change or vary the amount of the factors involved. Thus, if we vary amount of a certain hormone and find that a given factor changes in *amount* or *magnitude*, the factor in question can be given causal status in



rived, major research into the variables affecting the factors can be undertaken.

...the rough factorial map of a new domain will enable us to proceed beyond the exploratory factorial stage to the more direct forms of psychological experimentation in the laboratory (23, p. 56).

But, the facts of the case are that this more direct psychological experimentation has been largely programmatic. Only recently have we had large scale attempts to get "behind" the factors and find out what variables influence them. For example, Cattell (4) is undertaking studies on 1,000 children in an effort to determine the influence of genetic and environmental factors on the capacities of these subjects. Such studies as this must be carried out if factor analysis is to reach a stature whereby it allows an understanding of behavior over and above economical description. Thus, factor analysis first defines the important responses by cross-sectional analysis and then subsequent longitudinal studies may aim at discovering causal factors affecting these responses. No one is going to assert that those using factor analysis are laggards; Cattell reports that some 4,000 tests have been used to explore personality (4).

Although I might seem somewhat critical of factor analysis as practiced thus far, I would quickly add that those of us not primarily interested in it as a technique might make better use of it than we do. For example in the field of learning, even a restricted area such as rote learning, we do not know the relationships among performances on various tasks. If we could have a grand factor analysis of a large number of rote learning tasks we could define the essential skills involved. Furthermore, we could then choose tasks which are relatively pure on these skills for subsequent experimentation. And then, if we manipulate a given condition for these representative tasks we can make statements about the universality or lack of universality of the resulting relationship. At the present time, for example, if we determine the influence of a variable on paired-associate learning we haven't any sound basis for generalizing, say, to maze learning, for we do not know how the skills necessary for these two tasks are related. In short, factor analysis, in my opinion, still has a large part to play in many areas of research, not simply in determining personality traits.

crass a fashion. That is, usually there will be at least a crude hypothesis which determines the choice of the stimulus condition to vary for a particular experiment. Nevertheless, the question describes the essential condition of research which is stimulus-oriented.

In my way of thinking, low-level cause-and-effect relationships appear with sharpness when we consider stimulus-oriented research. If we run a carefully controlled conditioning study in which the intensity of the conditioned stimulus is systematically varied, and if we find a related change in acquisition of the conditioned response, I shall say that changes in the intensity of the conditioned stimulus caused the change in behavior. If one wishes he might think of this as apparent cause, thus making explicit the recognition that there are mediating mechanisms (physiological mechanisms) which are the more immediate cause. Indeed, if one wishes to pursue the matter further he can reduce it to specific nervous function, or to chemical reactions, or whatever level of explanation one desires and is capable of justifying. Or, certain theorists may postulate hypothetical processes which are related to the stimulus manipulation and these may be thought of as the cause. But, at the sheer empirical level, at the level of analysis of the experiment, the manipulated condition is as true a cause as we can possibly have. Empirical laws between stimulus variables and response measures are the basic facts from which more elaborate cause-and-effect chains start.

I would like to make two other preliminary comments. First, while the problem of experimental design will be taken up in later chapters, I think it well to note the basic design problem present in all stimulus-oriented research. Obviously, if we are going to vary a given stimulus condition, and observe changes in behavior, the essential dictum is that only one such condition (be it a very simple or a very complex condition) should be allowed to vary systematically. This is commonly said to be holding all conditions constant except one. Some comments have appeared in recent literature which imply that this basic principle of experimental procedure is outmoded. This is not true. One may vary more than one stimulus condition in a given experiment (multivariate designs) and it is very efficient to do so. But to draw a conclusion about the influence of any given variable, that variable must have been systematically manipulated alone somewhere in the design. Nothing in analysis of variance, co-

the same sense that any other construct is given such status. While it might seem redundant to think of both the hormone and the factor as causing the change in behavior in a one-to-one relationship, we shall see in later chapters that many psychologists find it convenient to do so.

I think I must make it clear that I am not saying that correlation never means causality; I am simply saying that we must be very thoughtful about the matter before reaching such a conclusion. I think it is fair to say, albeit a relatively meaningless statement, that response correlations allow us (if we wish) to infer some common causal condition (process, state, capacity). It is meaningless because we can infer the existence of such a hypothetical process with only a single response measure; we don't need a correlation. The situation is not amenable to cause-and-effect analysis until we can show how certain independent conditions will change the amount of a given factor as inferred from changes in scores on tests from which in turn the factor was inferred.

In actual practice such inferences are usually made following some form of stimulus analysis and manipulation, a topic to which I now turn.

### STIMULUS ANALYSIS

We have noted above that the essential rationale of factor analysis is to infer certain basic capacities of the organism as a result of response correlations. The stimuli involved in this situation are the tests—the original battery of tests, the scores on which provide the raw data from which in turn the factors are extracted. These tests are selected because it is believed by the investigator that they will tap all basic capacities in a given domain. In the usual sense of the word, there is no single continuum along which the tests are ordered before testing begins. The research is clearly response-oriented. Stimulus-oriented research, on the other hand, has as its basic premise the manipulation of a specified stimulus characteristic and the determination of change in behavior associated with the change in the stimulus. Put simply, the question is asked: "What variable stimulus conditions, when filtered through the organism, produce systematic changes in behavior?" This, of course, is the baldest type of empirical question, and probably very few research workers operate in so

a particular feature of the environment; the investigator does not choose one randomly. Several illustrations will show that this is a most common type of research.

1. Intelligibility of speech as a function of background noise. In such a study the investigator seeks a relationship between the intelligibility of transmitted speech and the intensity or complexity of background noise.

2. Alpha rhythm as a function of flash duration. Here the experimenter exposes the subject's eye to varying durations of light flashes and measures the attendant alpha rhythm of the occipital lobe.

3. Forgetting as a function of length of retention interval. Variations in time is one of the most common stimulus manipulations and occurs in nearly every area of psychological investigation.

4. Reading speed as a function of intensity of illumination. Work output as a function of type of music being played in the factory. Intelligence quotient as a function of nature of early environment. And on, and on. There is almost no end to the number of potential variables which constitute our environment. I want to consider just one more case, which is experimentally the same type of bald relationships suggested above, but which conceptually should perhaps be kept distinct.

5. Variations or manipulations in features of the environment are often conceptually related to changes in hypothetical processes in the organism. Variations which are produced in motivation by varying the amount of reward provide an illustration. Here the investigator, by manipulating the amount of food or amount of money, produces, or hopes to produce, changes in a process or state which he calls motivation. These changes may in turn be related to performance on a standard task. Now actually, any of the previously given illustrations could be conceptualized in the same way if the experimenter were so inclined. Thus, we might postulate neural blockage of some kind as being the intermediary between differences in flash duration and the alpha rhythm. It should be clear, therefore, that the operations involved in all of the above cases are basically the same regardless of how the experimenter may conceptualize his particular problem. I merely note this issue here for I will return to a full consideration of it in later chapters.

variance, latin squares, Greco-Latin squares, or Greco-Arabic-Latin squares has abrogated the basic principle. These powerful designs and statistical tools may save wounded experiments, and they provide remarkable levers for extracting variances, but in actual operation there are no laws resulting from their use which obviate the necessity of holding all factors constant except one if we expect to conclude anything about the effects of the factor.

A second preliminary point concerns the handling of variables whose influence is unknown. Suppose we want to study the effects of variable A; what do we do about potential variables B, C, D, etc.? There have been statements pertaining to this situation which have a certain amount of nonsense in them:

...the precise testing of a hypothesis generally presumes that one knows the relevant variables in the area of investigation, since without this knowledge it becomes difficult to establish adequate experimental controls. In such a case, an exploratory or formulative study is more likely to be fruitful than an experimental study (15, p. 29).

If this statement is taken at face value we would still be doing exploratory or formulative studies in all areas, for who can say when we know what all relevant variables are. Sound and precise experimental research does not hinge on our knowing what the relevant variables are; for the moment I ask you to accept this statement, as its defense and elaboration will not come up until later.

The following discussion treats of two kinds of stimulus analysis, one in which there is active stimulus manipulation and one in which natural variation occurs and conclusions are drawn on the basis of post-hoc statistical control. The first will be broken into three sections depending on the nature of the variables being manipulated.

#### ACTIVE STIMULUS MANIPULATION

*Environmental variables.* In these experiments the investigator chooses some feature of the environment which is capable of some form of quantification (as discussed in connection with response measurement) and his conditions consist of different amounts (perhaps only "qualitative" differences) of this feature. As mentioned previously, past results or theory usually dictate the investigation of

task that was being tried out for subsequent experimental use. These two types of instructions might be expected to produce differences in performance on the task.

2. In a rote-learning task, what is the influence of instructing one group of subjects to guess and another group not to guess?

3. In studying performance on a paper-and-pencil test, what is the influence of a "speed set" *versus* an "accuracy set?"

4. In a social-psychology experiment, what is the influence of telling one group of freshmen that their leader is a senior while another group is told that the same person is an instructor in the department?

#### NATURAL VARIATION WITH STATISTICAL CONTROL

In this method of research, no active stimulus manipulation is involved. Record keeping is a fetish in our social order, both in governmental institutions and private institutions. It, therefore, becomes theoretically possible to go back to records of individuals and try to find factors which are related to differences in behavior which have been noted. There are at least two different ways by which this has been worked out. First, different individuals may have actually been treated differently in some way. The investigator now attempts to search the records to see if the behavior differed as a consequence of the treatments. Secondly, differences in behavior of individuals may have been noted, and the investigator goes back to the records to see if there is one or more factors which might account for the differences in behavior. Again, let us look at some illustrations.

1. For many years at the University of Minnesota a student counseling program has been carried on. Eventually someone began to wonder if this program was worthwhile. In an attempt to answer this question the investigator went back to the records, took a group that had been counseled and a group which had not been counseled and then made comparisons of subsequent scholastic achievement.

2. In an investigation of a sociological nature (5) the investigator tried to answer the question as to whether or not participation in a Boy Scout program contributed to better community adjustment. His procedure was to go back to available records, obtain a group that had had several years of scout work and a group that had had

*Task variables.* In all of the above illustrations the manipulated condition was extrinsic to the particular task on which the subject was measured. In the case of task variables, some particular characteristic of the task itself is varied and resulting changes in behavior noted. Again, a number of illustrations seem to be the best way to give a picture of this kind of stimulus manipulation.

1. Rorschach responses as a function of the color or shading of the blots. Here, the actual task eliciting the behavior is changed in a certain way and observations are made of the differences in behavior which result.

2. Rote learning as a function of meaningfulness of the material. The classical illustration is variation along a scale of meaningfulness of nonsense syllables. Does rate of acquisition of the syllables vary as a function of different levels of meaningfulness?

3. Rate of acquisition of the pursuit-rotor skill as related to speed of rotation of the target.

4. Test performance as a function of the number of alternative choices allowed on a multiple-choice type quiz. Or, test performance as a function of similarity of the various wrong choices to the correct choice.

*Instructional variables.* While this form of stimulus manipulation might possibly be conceptualized as either environmental or task manipulation, I think it best to list it independently. In this type of research we attempt to vary the behavior of the subject by varying what we tell him about the task he is going to work on, or what the implication of his performance is, or how he should attack the problem, and so on. There is almost no limit to the number of possible variations in instructions, although in actual practice not a great many have been investigated. In general, the intent of varying instructions is to change the subject's perception or evaluation of the situation and to determine whether or not such changes are related to his performance. One might also prefer to list this type of experimentation under *subject manipulation* (to be discussed shortly); but, let us look at some illustrations to see the nature of the manipulation without too much concern for the niceties of classification.

1. Learning as a function of ego-involvement. We might give one group of subjects a learning task and tell them this task is actually a measure of intelligence, another group being told that it is a simple

However, it is when a factor responsible for the differences is supposed to have been found that one must look at the procedure with great care before accepting the findings.

#### QUANTIFICATION OF STIMULUS DIMENSIONS

The discussion under this heading can be quite brief since the points that we have made concerning quantification of responses in general apply here also. We have noted that the more precise the quantification of the response, the more precise the prediction which can be made. In the same fashion, the more precise the quantification of our stimulus dimension, the greater the precision in our laws resulting from their manipulation.

The stimulus dimension may be described along a physical scale or along a psychological scale. We have discussed these scales previously and have seen how the construction of psychological scales for characteristics of objects or symbols often precedes actual stimulus manipulation of the characteristic. I cannot emphasize too strongly the importance of the classical psychophysical methods (and derivatives from them) as techniques for dimensionalizing stimulus characteristics. Paired-comparisons, rank order, single stimuli, or even the methods of constant stimuli and average error can be adapted to these problems. All are powerful and extraordinarily useful tools for dimensionalizing characteristics of materials for which no physical scale is appropriate. All psychological dimensions used as stimulus variables eventuate from the reliable scaling of response.

Now, while we must continue to hold up precision of measurement as a goal toward which we continually work for all of our stimulus and response variables, we must also keep clearly in mind the fact that research with a certain limited usefulness can be done with extremely crude quantification of the stimulus. In the first place we may have coarse stimulus dimensions in which quantitative differences are expressed entirely in terms of words. For example, suppose we wanted to measure Thematic Apperception Test responses as a function of amount of trauma depicted by the cards. Judges might sort the cards into three piles representing high, medium, and low trauma, and, if this could be done reliably, we could proceed with the experiment. But we can be much more crude than



but little such work, and then measure differences in community adjustment.

3. At one time when I was in the Navy, and time was particularly heavy on my hands, I decided to test the hypothesis that boys raised on a farm had better mechanical aptitude than those raised in a city. Records were available which allowed me to test this hypothesis; we had records of whether or not the boys had lived on farms for an appreciable length of time and we also had their mechanical aptitude test scores.

4. In records of members of the armed forces we could find a group representing individuals who were discharged during the war for nonphysical reasons and a group which was not so discharged. We might then analyze other data in the records and see if we can isolate predischarge differences in these two groups.

5. We could divide married couples into two groups; those who have been divorced and those who have not been divorced. By going back into the history of these two types of cases we might discover a factor or factors which would seem to be related to the response measure (divorced or not divorced).

6. As a matter of fact, such problems can be worked out in the more staid experimental situation. One could keep records of various personal attributes of subjects as they serve in an experimental situation. We might then at a later time attempt to discover if any of these factors appear to be related to responses recorded in the experimental situation.

I think it can be seen that the number of problems which might be approached by this method is nearly inexhaustible. It is a fact, however, that not a great deal of research of this kind is undertaken, probably because some of the "design" problems involved are nearly insurmountable. There are published reports of research using these techniques which are utter farces as far as methodology is concerned. I shall later expose you to details of some of these investigations so that you can evaluate them for yourself. Nevertheless, the rationale of these studies is the same as a cross-sectional type of experiment using active stimulus manipulation. The basic idea is to have some measure of behavior, and then try to narrow causative factors down to one. Failure to find a factor which will account for observed differences is not serious and at least in a negative sense worthwhile.

is shown that with a particular age group learning is more rapid for material presented visually than for material presented aurally, training situations could well make use of this fact. But, it should be clear that from the analytical view of science as presented here, showing such a difference would merely establish a phenomenon on which further analytical work is required in order to derive specific causal relationships.

#### UNITARY AND COMPLEX DIMENSIONS

Stimulus dimensions may be unitary or they may be complex. By a unitary dimension I mean one in which only a single discernible characteristic is reflected. By a complex dimension I mean one in which two or more unitary dimensions combine to form one descriptive dimension. I do not wish to restrict complex dimensions to two levels, for, as we shall see, several combining levels may be apparent.

Our simple physical scales are intended to measure relatively unitary characteristics. Frequency of sound wave will be said to be unitary as measured in cycles per second. Many, if not most, psychological scales are complex. This is said with some evasiveness for a reason which will become apparent shortly. The dimensionalized characteristic is constituted of subsidiary dimensions which combine to make up the characteristic actually scaled.

I have said that the success of dimensionalizing (with other than physical scale) of any characteristic of behavior or the characteristics of objects rests on the reliability of the human discriminatory response. That is, to become repugnantly repetitious, we must have reliability in our measuring instrument. Now, it is quite possible to scale a complex dimension reliably. I suspect that any attitude that is scaled represents the composite of several subsidiary dimensions. For example, attitude toward socialized medicine could be dimensionalized along a single complex dimension. This dimension results from some sort of summation of subsidiary dimensions, such as, say, attitude toward bureaucracy in general, state of health, financial status, and so on. It seems evident that in order to scale a complex dimension reliably, the relevant subdimensions must vary in some systematic fashion with each other. There are a number of reasons

this. Let us assume that we wanted to find out whether auditory presentation resulted in faster learning than visual presentation of material. We could clearly specify two such different conditions and the research could be carried out simply, but about all we could say when we finished is that there are differences or no differences in learning as a function of mode of presentation. We could not, with any assurance, state on what particular dimensions auditory and visual presentation differ; one simply uses the visual system and the other the auditory system.

Let us look at another illustration. In a market research study we might want to determine which type of furniture, modern or traditional, was more preferred by a representative group of housewives. Probably, we could discover that there are very strong biases on such matters, but to relate these biases to the particular characteristics common to both traditional and modern furniture would be an extremely tedious, and perhaps impossible, job. The major implications of these illustrations is that they make apparent that in such crude research the ability to infer cause-effect relationship is seriously restricted, for we cannot state on common dimensions all the ways in which such complex stimuli differ. This problem, that is, the problem of what may be called unitary *versus* complex dimensions, is an important one to which I will give more attention shortly. First, however, I wish to make two other comments concerning research in which stimulus differences are qualitative.

In research dealing with these qualitative differences the design of the experiment can be perfectly sound; the limitation lies in the nature of the question such experiments can answer. Specific cause-effect statements cannot be made in the sense that we have discussed these statements earlier. Yet, such research may have considerable value. One of the initial tasks of a science is to establish reliable phenomena with which to deal. These experiments in which qualitative differences are used may at least establish whether or not there is a phenomenon. If so established, subsequent analytical research can be undertaken in an effort to discover the particular dimension or dimensions in the stimulus condition which are responsible for the phenomenon. So, from a strict scientific point of view, such experiments may have at best only a mapping function. On the other hand, these experiments may have considerable practical value. Thus, if it

present the details of the procedures used, but those interested may refer to several contemporary papers (1, 18, 20).

To remove the above discussion somewhat from the abstract level, I will discuss an illustration which brings the issues down to a research level. This illustration actually involves characteristics of the subject *per se*, a topic to be considered in the next section. I will anticipate this section in this illustration because it is especially suited to emphasize the points I am making.

Let us imagine that we wanted to carry out an experiment on the relationship between degree of adjustment of college students and critical flicker fusion frequency. I feel quite sure that able diagnosticians (whether psychiatrists or psychologists) could "sort" a large group of college students into a minimum of three groups, one being a poorly-adjusted category, another well-adjusted, and the third in-between. As usual, we would insist on reliability of the sorting. Having done this, we have scaled a dimension of adjustment. But now, look at the characteristics of behavior which must have been considered by our judges in arriving at a decision for the placement of an individual in one of the three categories. One major dimension which a judge might use could be called "amount of anxiety." But, the judge doesn't observe anxiety directly; rather, he observes other lower-order characteristics in order to make a judgment about amount of anxiety. For example, he might inquire into frequency of stomach disorders, dream content, fingernail biting, and other behaviors which he believes to be indicators of anxiety. The amount of each of these would then be "summed" to get a judgment of anxiety. Then, there would be other major subdimensions of degree of adjustment, such as amount of withdrawal. A sum of these major subdimensions, the amount of which was in turn determined by summing more unitary dimensions, becomes his final index of degree of adjustment. The diagnostician may, of course, use certain tests to aid in establishing the amount of a given characteristic, but it is quite clear that a great deal of a kind of mental "factor analysis" is involved in considering the importance of certain characteristics for the total picture, how the characteristics interact, how they combine, how much various ones should be weighted, and so on. When such multivariate mental manipulation and mensuration must take place one can but wonder why diagnostic attempts have any reli-

(which we need not consider here) why we might fail to dimensionalize reliably a complex dimension, but when we do dimensionalize one reliably we must infer that there is a systematic relationship among the relevant subdimensions.

I bring up this rather difficult problem of unitary and complex dimensions because one of the major tasks of our science is to reduce complex dimensions to their relevant unitary components. By this I mean the determination independently of the unitary dimensions which combine in some fashion to produce the complex dimension. It is only in this way that we can derive our most scientifically useful cause-and-effect relationships, namely, the relationships between dimensions described by a relatively unitary scale and the behavior which results from the manipulation of that dimension. Only in such instances do our laws become what I have called precise. And we will, at the same time probably discover that characteristics which we believed to be significant contributors to the complex dimension were in fact not.

I do not think that any of us fail to see how complex dimensions may well be constituted of subsidiary dimensions. Yet, as these complex dimensions become broken down into more and more subsidiary dimensions, a real question may arise as to how we can tell when we have arrived at a relatively unitary dimension of behavior. I know of no satisfactory answer to this question from a practical research point of view. Before any complex dimension can be broken down, the investigator must have ideas or hypotheses concerning the nature of the subsidiary dimensions so that some sort of independent scaling attempts can be undertaken. Ideally, when a complex dimension is reduced to a set of subdimensions which are relatively unitary, these subdimensions may become the manipulable stimulus conditions. Each may be manipulated independently to evaluate its influence, if any, on behavior which the investigator believes relevant.

But, it is only recently that systematic attempts have been made to break down complex psychological dimensions into their component dimensions. In general, some form of factor analysis or derivative therefrom is being used most successfully in this very important work. It is beyond the scope of the present discussion to

outstanding contribution by reducing the large number of apparently diverse characteristics to a workable number of basic or fundamental characteristics.

3. Characteristics of subjects may have varying degrees of stability or permanence. To identify extremes again, race is a very stable characteristic, intelligence somewhat less stable, and so on down to extremely transient characteristics such as sunburn or anger. However, most research on subject variables deals with the more stable characteristics.

4. Dimensionalized characteristics may be complex or unitary.

To return now to the question posed earlier, we see that there are two different kinds of research situations aimed at discovering relationships pertaining to subject variables. First, we might ask whether a specific subject variable is a relevant one for a particular performance. For example, we may "manipulate" the subject variable of anxiety to discover whether or not negative transfer is related to amount of anxiety. Anxiety is manipulated by choosing groups of subjects showing reliably different amounts of this characteristic. Let us contrast this with the case where an environmental or task variable is manipulated. When we manipulate an environmental variable, our *groups of subjects* must not differ systematically on any subject variable. When we manipulate a subject variable the *environmental variables* must not differ systematically for any of our groups of subjects. (It might be parenthetically mentioned that in this particular illustration of anxiety we are not necessarily limited to selecting groups of subjects differing on anxiety. We might very well institute experimental conditions which would induce differing amounts of anxiety in groups originally no different on this characteristic. However, there are many subject variables in which this would not be feasible. It would be rather difficult to imagine how, in a short experimental period, we could induce varying amounts of intelligence among groups which initially did not differ on this factor.)

A second type of research problem would attempt to determine conditions influencing the subject's characteristics as such. What are the factors responsible for different amounts of intelligence? What are the factors responsible for schizophrenia? What are the factors

ability at all. It is just such situations as this which makes it all the more apparent that complex dimensions must be broken down into subdimensions which are more unitary in nature so that the influence or weight of each can be determined independently of the influence of others. I make no pretense that this can be done easily, but I also have made no assertions concerning the ease with which problems confronting a scientist are worked out.

### SUBJECT ANALYSIS

I have thus far discussed some issues which I have felt to be especially germane to response measurement and stimulus analysis. The third and central component in the research situation is the subject himself; indeed, the subject or organism is obviously the *raison d'être* for psychological research. Some stimulus-oriented researchers would sometimes almost appear to resent the fact that a subject is necessary in order to derive relationships between environmental (or task) variables and responses. But, whether we like it or not, the subject remains.

As indicated earlier, the number of environmental and task variables which might potentially influence behavior is almost unlimited. So also is the number of characteristics or variables of the subject. Any characteristic of the subject which might be shown to vary reliably in amount among subjects is potentially a relevant subject variable. But a variable for what? Let me back up a bit before this question is answered. Because the previous material in this chapter has direct bearing on the needed discussion at this point, a series of brief statements should bring us to a position where subject characteristics or variables can be placed in their proper perspective.

1. The first step in any such research is to quantify characteristics on which subjects differ. This quantification may take place at all levels and by all methods discussed earlier. We may use physical scales to obtain, for example, height, weight, and chronological age differences. At the other extreme we could identify qualitative differences such as race or occupational differences.

2. On the level of personality and intellectual differences alone it would seem that the number of characteristics on which subjects differ is hopelessly large. It is here where factor analysis makes an

4. CATTELL, R. B. *Personality: A systematic theoretical and factual study*. New York: McGraw-Hill, 1950.
5. CHAPIN, F. S. *Experimental designs in sociological research*. New York: Harper, 1947.
6. COOMBS, C. H., RAIFFA, H., & THRALL, R. M. Some views on mathematical models and measurement theory. *Psychol. Rev.*, 1954, 61, 132-144.
7. DOLLARD, J., & MOWRER, O. H. A method of measuring tension in written documents. *J. abnorm. soc. Psychol.*, 1947, 42, 3-32.
8. DUNCAN, C. P. Transfer in motor learning as a function of degree of first-task learning and inter-task similarity. *J. exp. Psychol.*, 1953, 45, 1-11.
9. ELKIN, A. *Personality as a variable in serial verbal learning*. Ph.D. dissertation, Northwestern Univer., 1950.
10. EYSENCK, H. J. The logical basis of factor analysis. *Amer. Psychol.*, 1953, 8, 105-114.
11. GUTHRIE, E. R. Psychological facts and psychological theory. *Psychol. Bull.*, 1946, 43, 1-20.
12. HALL, J. F., & KOBRICK, J. L. The relationships among three measures of response strength. *J. comp. physiol. Psychol.*, 1952, 45, 280-282.
13. HARLOW, H. F., & MEYER, D. R. Paired-comparisons scales for monkey rewards. *J. comp. physiol. Psychol.*, 1952, 45, 73-79.
14. HEYNS, R. W., & LIPPITT, R. Systematic observational techniques. In G. LINDZEY (Ed.) *Handbook of social psychology*. Vol. I. Cambridge: Addison-Wesley, 1954.
15. JAHOOA, M., DEUTSCH, M., & COOK, S. W. *Research methods in social relations*, Part I. New York: Dryden, 1951.
16. MARQUIS, D. P. A study of frustration in newborn infants. *J. exp. Psychol.*, 1943, 32, 123-138.
17. MEADOW, A., GREENBLATT, M., LEVINE, J., & SOLOMON, H. C. The discomfort-relief quotient as a measure of tension and adjustment. *J. abnorm. soc. Psychol.*, 1952, 47, 658-661.
18. OSGOOD, C. E., & SUCI, G. J. Factor analysis of meaning. *J. exp. Psychol.*, 1955, 50, 325-338.
19. SCHEIDT, H. *Individual meaningfulness ratings and speed of learning with observations on retroactive and proactive inhibition*. Ph.D. dissertation, Northwestern Univer., 1954.
20. SHEPARD, R. N. *Stimulus and response generalization during paired-associate learning*. Ph.D. dissertation, Yale Univer., 1955.
21. SINGER, P. *The relationship between childhood punishment and overt*



causing differing thresholds for embarrassment? In this kind of research we start initially with subjects *not* differing on the subject variable under investigation and then vary environmental conditions in some systematic fashion to see if the subject characteristic is affected. In the parenthetical comments above where conditions were varied to change anxiety we were doing this very thing. The well-known identical twin studies are essentially of this nature. At birth, identical twins are assumed to be equal on subject characteristics. If one of the twins is placed in one environment and the other in quite a diverse kind, differences with respect to, say, intelligence which develop may be attributed to differences in the environment. Research using natural variation and statistical control is sometimes employed to determine factors influencing subject variables. The illustration given earlier in the chapter in which the investigator attempted to determine the influence of scouting on adjustment is such a study.

Problems of research design in manipulating subject variables are theoretically no different from those present when stimulus manipulation is carried out. However, in actual practice, the working out of a design in which cause-effect relationships can be stated with confidence is a serious matter. With the rapid increase in the number of students in clinical psychology there has been a commensurate growth in research in which subject variables are manipulated. A very common procedure is to choose groups falling into different clinical diagnostic categories, expose all to a standard stimulus situation (such as a learning task) and observe what if any differences in behavior occur. If differences do occur they are attributed to differences in diagnostic categories. We shall see in later chapters that such experiments have often violated fundamental rules of scientific method.

## REFERENCES

1. ABELSON, R. P. A technique and a model for multidimensional attitude scaling. *Publ. Opin. Quart.*, 1955, 18, 405-418.
2. ALLPORT, G. W. Scientific models and human morals. *Psychol. Rev.*, 1947, 54, 182-192.
3. BROGDEN, W. J. Acquisition and extinction of a conditioned avoidance response in dogs. *J. comp. physiol. Psychol.*, 1949, 42, 296-302.

## Operational Definitions

When a writer says that terms should be operationally defined, or says that this particular concept is a poor one because it is *not* operationally defined, most psychologists have vague feelings of approbation. Surrounding the idea of operational definitions there is a general aura of righteousness that is at once dogmatic and scientific. Yet, I have a strong suspicion that if most of us were forced to defend operational definitions as an integral operating base of a science, we would be limited to mouthing a few clichés about what operational definitions are and what their value is.

Critical and general expository essays concerning operational definitions in psychology have appeared periodically since the idea of such definitions was formally introduced to psychology in the early nineteen thirties. The last major discussions primarily by psychologists appeared in 1945 when a complete issue of the *Psychological Review* was dedicated to the matter, and one by nonpsychologists in 1954 in the *Scientific Monthly*. Yet, from the students' point of view, there is no source which is at present adequate in its discussion of operational definitions. Too many of the discussions have been overlaid with a confusing shadow cast by philosophical clouds, obscuring the more pragmatic problems involved. I know of no source to which one can turn which actually evaluates detailed matters involved in constructing an operational definition. I think we have placed too much faith in gross transfer. We tell the student that an operational definition is one which specifies the meaning of the concept by denoting the measuring operations, and then we seem to expect him to formulate definitions which carry the essential operations. It is my experience that most need further guidance on this matter.

Do I accept a practice of operationally defining concepts? It is

- error frequency in serial adjective learning. M.A. thesis, Northwestern Univer., 1954.
22. STEVENS, S. S. Mathematics, measurement, and psychophysics. In S. S. STEVENS (Ed.) *Handbook of experimental psychology*. New York: Wiley, 1951.
  23. THURSTONE, L. L. *Multiple-factor analysis: A development and expansion of the vectors of mind*. Univer. Chicago Press, 1947.
  24. ZANDER, A. Systematic observation of small face-to-face groups. In M. JAHODA, M. DEUTSCH & S. W. COOK (Eds.) *Research methods in social relations, Part 2: Selected techniques*. New York: Dryden, 1951.

must be operationally defined. For the present chapter I am limiting the discussion entirely to the operational definition of natural phenomena. The above mentioned issues or problems will receive some attention in later chapters.

### PURPOSE OF OPERATIONAL DEFINITIONS

Having set myself firmly on the side of operationism, along with most psychologists, you may ask: "so what?" Of what significance is an alliance of this kind, and why be so ardent about the union? Perhaps I cannot answer such questions satisfactorily, for the written word has a certain bleakness or lack of expressiveness which does not entirely convey the convictions one develops over a period of several years of reading, talking, and composing critiques of scientific endeavors. For what they are worth, however, I will indicate three primary benefits which accrue to a science as a result of adherence to operationism. The remainder of the chapter may then be thought of in part as illustration of these points.

1. First, I would say that operational thinking makes better scientists. The operationist is forced to remove the fuzz from his empirical concepts. Certain kinds of operational definitions sharpen experimental designs. When the investigator is forced to ask himself whether or not his particular operations will allow him to derive a unique phenomenon (in the sense that it is differentiated from other phenomena), he may find himself adding certain control conditions which will produce the uniqueness.

2. Secondly, I must insist that a research-oriented operationism restricts the number of concepts of a science. New concepts are not introduced unless the operations which define the phenomenon clearly differentiate it from other phenomena. Certain logical and philosophical considerations of operationism (e.g., 3, 14) would lead one to expect that insistence upon operational definitions would expand the number of concepts in a science almost without limit. I do not think this has happened, nor do I think it will happen.

3. Finally, operationism facilitates communication among scientists because the meaning of concepts so defined is not easily subject to misinterpretation.

not a mere matter of accepting or rejecting operationism, for to ask such a question is to ask whether one accepts science as a technique for understanding the laws of nature. Indeed, I would assert that a criterion of whether or not a so-called empirical concept is a scientific concept is whether or not it has been operationally defined. The starting point of any science is, basically, a set of phenomena which have been operationally defined. For, in their simplest and basic form operational definitions specify the measuring operations used to identify phenomena. Thus, it seems to me that operational definitions, stripped of their excesses, reflect measurements, and reliable measurements are the roots of any science. Reliable measurements initially identify the phenomena with which a science concerns itself. An operational definition need not explicate a functional relationship; it may only reflect a demonstration that a reliable phenomenon exists. Further research may then be undertaken to try to learn more about the phenomenon by discovering of what variables it is a function and how it is related to other phenomena. Operational definitions are not a science for they need not express relationships and they are not theory, but they are the necessary base for a science.

An operational definition does not tell us much about what the important or relevant variables of the defined phenomenon are. To repeat, basically an operational definition simply tells us that there is a phenomenon. Bergmann and Spence point up this matter as follows:

We see that even at the level of the empirical laws the scientist cannot derive any help from operationism. He will have to rely upon his own ingenuity and whatever help he might be able to get from an articulate theory (2, p. 5).

I have said above that operational definitions are not theory. Yet there are certain issues regarding the relationships between what I have called here operationally defined concepts and theoretical concepts that must be explored. That is, operationally defined concepts are sometimes used as explanatory concepts; operationally defined concepts may, of course, enter into theoretical statements; operationally defined concepts are used by some writers as theoretical concepts. And we are told by some that theoretical concepts

ception to the specific statements on the ethnocentric scale is not clear, as it rarely is, for this translation usually remains private. We can only assume that the questions making up the scale were such that Levinson felt they tapped the process or state he wished to measure. If his scale is reliable, he has operationally defined ethnocentrism with it. Actually, the scale may not at all measure the process which he wished to measure; it may measure something quite different. Nevertheless, it is operationally sound. The danger involved is that the investigator will fall into the trap of thinking that because he went from an artistic or literary conception of ethnocentrism to the construction of items for a scale to measure it he has validated his artistic conception.

Therefore, we must guard against thinking that our literary statements about a phenomenon which we believe exists, and which we wish to measure, are inevitably identified with the phenomenon we finally measure. Literary definitions have their place in the communication scheme of science. In many cases I think they are very useful in getting the reader into the writer's frame of reference so that the operational definition does not come as a shock. But, these literary definitions are not accepted in lieu of, but only as an introduction to, the operational definition. In the formal sense of science, the operational definition alone must bear the name of the phenomenon.

In another study Bousfield (5) wished to test certain hypotheses about the relationship between the mood of a subject and the nature of verbal associates. The literary definition of mood is:

Mood is a generic term denoting general feeling tone which is a resultant of the specific feeling tones associated with the specific motivational systems operating at any one time (5, p. 67).

The operational definition of mood was very simple: subjects were told to: "rate your mood, or how well you are feeling at the present time" on a 10-point scale ranging from "feel as badly as possible" to "feel as good as possible" (5, p. 73). Now of course the subjects didn't know about Bousfield's literary definition of mood, and the operational definition does nothing by way of validating it.

*Infinite regress: levels of generality of operational definitions.* One

## LATITUDE ALLOWED

Partially as a means of setting before you illustrations of operational and nonoperational definitions, and partially as a means of outlining the scope of operationism, I will take up three general problems which inevitably arise in thinking about the topic. In subsequent sections of the chapter I will give extended illustrations of operational definitions and in conjunction with these illustrations make further observations of the value and limitations of such definitions.

*Literary lead-ins.* Definitions which are commonly found in standard dictionaries I will call literary definitions. At this point I wish to contrast literary and operational definitions but at the same time point out the usefulness that some of these literary definitions have in psychology. Sometimes improved communication may result if the scientist precedes his operational definition of a new concept by a literary definition. These literary definitions may help the reader understand the general nature of the phenomenon which the scientist is trying to bring under careful scrutiny. An illustration will show what I mean.

Levinson (7) was concerned with a subject trait which he called *ethnocentrism*:

Ethnocentrism is conceived here as an ideology: *a relatively organized, relatively stable system of opinions, attitudes, and values.* The term "opinions" refers to ideas about the nature of social reality; these include specific "factual" beliefs as well as more underlying imagery of groups and institutions.... They are the psychological facts and assumptions in the individual's conception of society. The term "attitudes" as used here refers to one's readiness for action; it includes all ideas about what should be done to, for, or against any social entity. Values are the individual's standards of right and wrong, good and bad. (7, p. 19).

At least in a general way, the above statements give one an appreciation of the behavior with which the investigator hoped to work. It is seen that the total state of ethnocentrism is believed to be a composite of three subsidiary states. Now, in this particular illustration, an operational definition of ethnocentrism was provided by a scale which Levinson constructed. Just how he got from the literary con-

The UR (unsuccessful reader) was defined as a child between the ages of 8-0 and 16-11 who had achieved *either* a Verbal or Performance Scale IQ of 90 or higher, who had fallen 25 per cent or more below the mean reading grade level on the Wide Range Achievement Test, for a child of his chronological age, and who had attended public or private school for the expected number of years for his given age (6, p. 268).

I would prefer to give first a general operational definition of the unsuccessful reader, perhaps as follows: "An unsuccessful reader is one whose reading performance falls a specified amount below expected performance for his age, education, and intelligence level." Subsequently, this general definition can be reduced to a specific criterion imposed in this particular study. But, however one prefers to state the definitions, let it be clear that at one point or another in the exposition the basic operations must be stated, and if it is necessary to regress in order to establish communication it must be done. I would repeat, however, that I have been unable to find any illustrations in the literature where undue hardships have been caused by infinite regress.

*Provisional definition.* Many times a scientist sets out to demonstrate a new phenomenon. That is, because of certain observations or because of theory, the investigator believes there exists a phenomenon which has never been investigated scientifically. If he gives (as he must) an operational definition of this expected phenomenon before the operations are actually carried out, I shall call it a provisional definition or pre-research definition. It must be provisional because the investigator may find either that he cannot carry out the operations or if he can and does, no new phenomenon is discovered. But, what the investigator does is to say: "If I do this, and if such and such happens, then I define the phenomenon by these operations."

A considerable amount of research has been built around concepts which Freud advanced on the basis of relatively nonsystematic observations. This research, in the last analysis, is an attempt to give scientific status to the concepts; it is an attempt to give operational definitions so that further work can be done to demonstrate the conditions which cause the phenomena to vary, their interrelationships, and so on. Or to put it bluntly, these studies have asked—or should



of the criticisms sometimes directed at operationism it that a strict interpretation of it would mean that each critical word in an operational definition must itself be defined operationally, each critical word in these definitions so defined, and so on, so that there is an infinite regress of definitions. Let me first concede that indeed such regression must take place if necessary to establish communication. But also let me quickly add that in actual practice the results of this necessary concession are rarely onerous. As a science develops, standard concepts are developed. By "standard" I mean concepts that are understood by all. (Such terms are sometimes called *primitive* terms.) Such concepts may then be used in new operational definitions without acquiescing to a demand for regress. For example, if the term "Stanford-Binet I Q" is used in an operational definition, few would need a second definition explicating what is meant by it.

Sometimes in defining a new concept rather omnibus operational definitions may result. Although these may seem oppressive from a belletristic point of view, if such length is necessary to expose the meaning of the concept, it must be accepted. Let me give you an illustration, which while long, is still incomplete:

....I shall define operationally the strength of the eathexis to any type of goal by the tangent of the angle made by the line resulting from plotting the magnitudes of the measures for getting-to such a goal against the magnitudes of the measures for getting-to the standard goal (13, p. 364).

This definition may need further elaboration by operationally defining what is meant by *goal*, *standard goal*, and *magnitude of the measure*. And in fact such definitions are given by the investigator.

If such omnibus definitions are necessary because of the lack of standard concepts, the investigator has two alternatives (to indicate the extremes). He may make only a general operational definition and indicate that the specific operations implied by the critical words in the definition will become clear when the details of the procedure are set forth. Or, the investigator may include in a single, long definition all detail that is necessary to establish communication. Personally, I prefer the former method as a means of keeping our scientific prose from becoming too ponderous. Here is an illustration of what I would call a ponderous operational definition.

discussed earlier. That is, I will not go into all the details of the operations for each illustration, but will assume you will accept the fact that the details of the operations could be added. Third, I have found it useful in some of the illustrations to lead up to the operational definition with an abbreviated literary definition.

The six approaches may be subsumed under two general headings, *response identification* (of phenomena) and *stimulus-response identification*.

### RESPONSE IDENTIFICATION

Three approaches fall under this heading. One I shall call *simple response identification*, one *complex response identification*, and the other, *scaling identification*.

*Simple response identification.* The purpose of the operations here may be twofold. First, and most frequently, the investigator exposes a group of individuals to a static set of conditions and the responses of the individuals to this set of conditions are measured. The operational definition is complete when it is shown that individuals do differ reliably, i.e., where it is shown that the rank orders of the scores remain relatively constant on a retest or by other acceptable techniques for determining reliability. Such definitions result from the initial standardization of any test, although they may be used in a variety of situations.

Secondly, there are some research efforts where the unit of measurement is a group, rather than individuals. In such situations the investigator must show only that groups differ reliably in their responses to a static set of conditions. Thus, the group score is the unit in the distribution. Certain social-psychology experiments make use of such definitions. Let us look at a number of illustrations for these two kinds of simple response identification.

1. Suppose you believe that there is a characteristic of behavior which no one else has defined and investigated. Let us call this hypothesized characteristic X. Your initial procedure is to construct a test, say, a paper-and-pencil test, which you think "gets at" this particular characteristic of behavior. You find that the test is reliable. Now your definition is simply: "X is the characteristic measured by this test," and you point out or exhibit the test. At this stage, of

have asked: "Is there a phenomenon to be operationally defined?" Can repression, displacement, or transference be defined operationally, or are these pseudophenomena resulting from the lack of control for Freud's observations or to false inferences from the observations? If the operations do establish that there is a certain phenomenon of the nature Freud reported, then an operational definition is given it and it can be brought into the realm of scientific investigation and discourse. If, on the other hand, no such phenomenon can be demonstrated by acceptable operations, the term remains outside the scope of science. This would not mean that Freud was wrong; it would only mean that the process he has indicated has no scientific meaning. If many and varied attempts are made to establish the existence of the phenomenon, and if they are all unsuccessful, the probabilities become great that Freud's observations, or his interpretation of those observations, were somehow in error.

### FORMULATION OF OPERATIONAL DEFINITIONS

For expository purposes I will say that operational definitions of phenomena are formulated by six different approaches. These six approaches are determined by the nature of the research situation. I make no claim that these encompass all research situations in psychology, nor that all can be clearly distinguished from the others. I do believe, however, that the discussion will cover most typical situations in which psychological research is carried on. To a large extent the six situations parallel the outline of the previous chapter where various components of the research situation were analyzed. I have named each of the six approaches, probably quite inadequately, but such naming at least allows us to have a breakdown of the material into rough classes and perhaps facilitates communication.

Three other prefatory comments are in order. As indicated earlier, the different approaches will be extensively illustrated and in conjunction with some of these illustrations I will bring up general issues about operational definitions which have not hitherto been discussed. I do this in this fashion simply because certain of the illustrations make very clear the issues involved. Secondly, I will keep my illustrations of operational definitions primarily on the general level, as

that it is not proper to ask such a question; it assumes that there is some other source to which one can appeal for truth. It is a perfectly valid question to ask: "Do you think that this scale measures anxiety in the full range of behavior to which the term is commonly applied?" Such a question can be answered by research if other or more expanded ideas of anxiety manifestations can be given adequate quantification. The facts are that manifest anxiety as defined by this scale has been found to relate to several other forms of behavior, i.e., it is a useful tool for further research and anyone who looks at the operational definition knows the meaning of the term *anxiety* as used.

4. Next, let us imagine that we wanted to give an operational definition of group morale, and that our groups are military units. According to certain considerations we might be led to believe that our literary conception of group morale would be manifested in different amounts by such behaviors as number of AWOL's, number of visits to the dispensary, number of letters written, and so on. Indeed, we might use a number of discrete measures and derive a composite of them for our index of morale. If we can show that the measures are reliable we have given an operational definition to morale.

5. The term *cohesiveness of a group* means, in my literary sense, the "togetherness" of people in a group, how loyal they are, how well they think as a group, and so on. A social psychologist might wish to go from such literary concepts to an operational definition. He might be led to conclude that the amount of such cohesiveness could be indexed by the number of "we's" occurring in the speech of the group in a specified situation. If he can show that groups differ reliably on this response index, we have given an acceptable operational definition to cohesiveness.

6. Intelligence may be defined as the score made or the characteristic measured by a specified test. It may be noted that different people think of the implications of operational definitions in two ways. Some think of the response measure as a base from which to infer a characteristic, state, or process of organism. Others may think of it on the strictly empirical level. Thus, in my operational definition of intelligence I have said "characteristic measured by" or "score." In a later chapter I will discuss in detail differences in

course, you have only defined your concept (X); what else it is related to, hence, how useful it is in understanding behavior, must be demonstrated by subsequent research. I shall examine this problem in more detail later when discussing another concept.

2. Next, let us suppose you construct a performance test which in your opinion is tapping skill in basic mechanics. If you show that this test measures reliably, you may then define Mechanical Aptitude as what is measured by performance on the test.

With regard to this test, let me set up an extreme situation. Let us say that you constructed a paper-and-pencil test which consisted largely of questions about the literature of the ancient Romans. Imagine further that the test measures reliably. Finally, suppose you say that mechanical aptitude is measured by performance on this test. This may sound ridiculous, and certainly, no one would ever do this. But, I wish to point out that your operational definition of mechanical aptitude, defined as performance on a test in which the questions are about Roman literature, is perfectly sound. That is, you can easily defend such a definition on the grounds of strict operationism. However, you cannot defend it on social scientific grounds; you certainly would be accused of a lack of propriety in assigning names. One of the purposes of operational definitions is to facilitate communication, and the above naming might actually hinder communication.

3. One of my colleagues (12) developed a scale to give operational definition to *manifest anxiety*. This scale, consisting of selected items from the Minnesota Multiphasic Personality Test which were judged by clinicians to "get at" anxiety, was shown to be reliable. So, the definition of anxiety is given by the measuring instrument; a person who makes a high score is said to have high manifest anxiety, a person obtaining a low score, low manifest anxiety.

Now, you may have an impulse, as a number of others have, to ask: "But does this scale *really* measure anxiety?" What this question seems to imply is that the one asking the question doubts that the scale measures all characteristics of behavior which have been labeled anxiety by clinicians. This may be quite true, but, such considerations are irrelevant when considering the adequacy of the definition  
; perfectly sound. So, I would say

range of behavior tested and which would be defined as intelligence. (Cf., Spiker & McCandless, 10, for an extended discussion of the status of the concept of *intelligence*.)

7. Simple response-defined concepts are not limited to testing situations in the clinical sense. In the field of learning, for example, there are many such definitions. *Operant level* may be defined as the number of unreinforced bar pressings in 30 minutes in a Skinner box; or, *exploratory behavior* as the number of traversals of 15-in. maze units per unit of time for rats satiated with water and food (9).

I think these are enough illustrations of operational definitions which I have tagged simple response definitions. These are the most elementary operational definitions possible in psychology. In the situations to which these are applied we need only demonstrate a reliable individual (or group) difference in behavior.

*Complex response identification.* Essentially, definitions falling in this category are elaborations of simple response identification. Several discrete response measures go together to define the concept. These response measures are not "put together" in an arbitrary fashion; rather they become the base for a concept only if certain correlational criteria are met. The best illustration of this type of definition is given by factor analysis. Let us review this procedure briefly.

A large group of carefully selected tests is given a sample of individuals. Each individual test, if reliable, can form the basis for an operational definition as discussed under simple response identification. By factor analysis, intercorrelations among test scores are determined and a group of tests which intercorrelate among themselves, but not with other groups, define the factor. Thus, the factor is not directly defined operationally in terms of its differentiating among individuals, but rather in terms of its uniqueness, i.e., it must have little or no relationship with other factors derived from the same battery of tests. So, the final definition is in terms of performance intercorrelations among a group of tests.

Whenever a certain pattern or constellation of responses is needed to differentiate one phenomenon from another we use complex response identification. Identification of clinical syndromes is another illustration. Presumably, no single characteristic of behavior will differentiate one clinical syndrome from all others. Clinical

conceptual levels of thinking. At the present time let me say only that both methods of thinking about such concepts are used. In the case of the concept of intelligence I want to again digress to discuss issues about the concept which periodically crop up in our literature.

For some insistent reason many psychologists, when considering the concept of intelligence, have developed the peculiar idea that not only should a definition define but it should lay bare the very nature, essence, or significance of intelligence. An operational definition is a definition; it is not supposed to expose the whole truth of nature or the order of the universe. In the case of intelligence all the definition does is to state that individuals' performances differ reliably on a test that is called an intelligence test. It has been said that an operational definition of intelligence such as I have given is as sterile as opposed to a "dynamic" concept. I am not concerned about the fertility of definitions nor the dynamism of words; I am concerned only with specifying the basic meaning of a concept. Response-identified operational definitions are not supposed to represent relationships between dependent and independent variables. Basically, an operational definition asserts only that a phenomenon has been reliably measured. When we operationally define other concepts, say learning, we are not expected to indicate in the definition all the variables of which the phenomenon is a function, how it will help raise the level of civilization, or how to make the United Nations an effective instrument for world peace. So too, when we define intelligence we are not obligated to show that it correlates with school grades, that business executives have more of it than do ditch diggers, or that geniuses have more than idiots. The operational definition exposes the operations by which you are quantifying the phenomenon; it is not both the starting and stopping point of science. There are those who complain that an operational definition of intelligence does not *really* say what intelligence is. To say what intelligence is in a scientific sense is to say what scores on the test relate to, and relationships are obtained by investigation and research, not by definition. Of course, it is always a perfectly legitimate question to ask whether or not a particular test samples well all the behavior which in a literary sense would be called intelligent behavior. Such a question might even lead to a broadening of the

general type of operations involved, a single illustration should be sufficient.

Let us imagine that an investigator sets out to dimensionalize the *affective tone* of verbal materials. He believes that certain words evoke feelings of unpleasantness, others feelings of pleasantness, others neutral feelings. To a group of judges he presents say, 100 words, which he thinks are representative of the dimension as he conceives it. The judges are asked to rate these words on a 7-point scale from most pleasant to most unpleasant. If he shows that the words (although not necessarily all of them) can be allocated along the scale in a reliable fashion by the judges, he has, by noting his procedures, given an operational definition to affective tone. He may further wish to specify, for example, that all words with a rating of 2.0 or less will be called unpleasant, all of those with a rating of 5.0 or more pleasant and all in between neutral. If he specifies the instructions to his judges, and other relevant details of the scaling technique, the operations are repeatable.

I would like to point out in concluding this section on response identification that none of the operations produces a law or relationship in which stimulus variables are involved. This is intimately related to the discussion in the previous chapter on cause-effect analysis. Response correlation rarely gives us a basis for inferring cause and effect. We shall see that in the next general type of operational definition, stimulus-response identification, we usually have at least a crude stimulus-response relationship indicated in our operational definition since some form of manipulation of a stimulus variable is necessary to define the concept.

And, let me repeat that the scientific usefulness of concepts defined by response identification is not indicated by the definition. What variables influence the phenomenon defined, how it is related to other concepts, and so on, are matters for subsequent research.

#### STIMULUS-RESPONSE IDENTIFICATION

Under this general heading I have three divisions indicating somewhat different sets of operations. Again, I am not completely satisfied with the names nor with exclusiveness of the categories, but they are the best I have been able to formulate.



groups are based on a composite of several responses. Insofar as such grouping can be reliably formed, whether by tests, clinical judgment, physiological measurements, or combinations of all, an operational definition is formulated provided the criteria for inclusion and exclusion in classes are specified. The response measure, regardless of level of quantification, must be a part of the public record in the definition. We cannot define a concept operationally on the basis of unquantified intuitions. The operations must be repeatable by others.

The use of factor analysis to define concepts has certain merits relevant to the over-all picture of operationism. I have said that one of the virtues of operationism is that it restricts the number of concepts. A concept, to be operationally defined, must result from procedures which are basically different from those used to define another concept. This problem concerning plurality of concepts will be returned to later in the chapter, but I need to open the discussion at this point. An operational definition of a phenomenon can be made by the use of a single, simple, short, paper-and-pencil test according to the criteria discussed under simple response identification. Now, conceivably, the number of such phenomena which could be so defined is almost unlimited. It therefore becomes apparent that the number of operationally defined concepts might be multiplied excessively. And it is a fact that we have no over-all plan in our scheme of science of psychology to avoid such multiplicity. But it is here where factor analysis makes a strong contribution. Independent operational definitions should be maintained only for unique phenomena. By factor analysis, tests which measure essentially the same characteristic of behavior lose their individual identity for definitional purposes. Each test becomes just one of a group of tests measuring the same characteristic, and a single definition is used to reflect the operations defining the factor measured by all the tests.

*Scaling identification.* As mentioned in the previous chapter, the human discriminatory response may be used to dimensionalize characteristics of objects or events (such as responses). The procedures involved constitute the operational definition of the characteristic. Since we have discussed at some length in the previous chapter the

varying amounts of the perimeters missing and asked subjects to report when they did and did not see a triangle. He then defined closure as the amount of perimeter which must be present before 50 per cent of the subjects reported seeing a triangle. The steps in this procedure operationally define closure.

*S-R identification with psychological scale.* We have seen how psychological scales are derived by the human discriminatory response. The operations involved define the dimension of behavior (or of objects) reflected by the scale. Now, however, when we use such scales as manipulable stimulus dimensions, and observe concomitant changes in behavior, we are in a position to give an operational definition of a phenomenon which makes use of the response-defined stimulus scale. However, in most instances when such operations have been used, the manipulation of the stimulus variable is thought of only as influencing another phenomenon. For example, manipulation of affective tone may influence rate of learning, but these operations are not the essential ones for defining learning. Nevertheless, when such stimulus manipulations are carried out and when they influence behavior, we have, in one manner of speaking, a new phenomenon. And, whether one wants to call it a phenomenon or not, it is a relationship which independently requires operational definition. For some reason, however, names have not been customarily assigned to such relationships. But, failure to name has nothing to do with defining. In other words, we can manipulate, say, meaningfulness, note its influence on rate of learning, and the relationship so derived is defined operationally by giving the operations involved.

Some investigators have attempted to demonstrate the phenomenon of *repression* by measuring the relative retention of pleasant and unpleasant words. A group of pleasant words and a group of unpleasant words are learned and retention is measured after a period of time. If the retention of the unpleasant words is poorer than the pleasant, repression is said to be defined. One could simply look at such operations as determining the influence of affective tone of words on retention; here, however, a name is assigned to the relationship. In this particular illustration we have at best had only a provisional operational definition; results of such (and similar) operations have been undefinitive in demonstrating a new phenomenon.

*S-R identification with physical scale.* By this technique the operational definition of a concept is obtained by showing a relationship between a physical scale of some sort (a physical stimulus scale) and the responses prompted by that scale.

How do we operationally define pitch? Basically, the operations require only that we vary cycles per second of a sine wave and that reliable judgments concerning variations in highness or lowness of the sound be reported. It so happens that in this case the amount of phenomenal highness is directly related to cycles per second. Of course, as discussed earlier, the complete definition would require specification of the particular technique used to present the stimuli, e.g., constant stimuli, and perhaps further elaboration of this technique if it deviates from the method as commonly understood.

Now, as we can see, such operations not only define pitch but may establish the empirical relationship between cycles per second and the phenomenal change in sound we call pitch. The definition not only identifies the critical variable needed to produce the phenomenon, but also gives us a lawful relationship. In the simplest case of such a definition we would present only two variations in cycles per second and if the judgments in sound changes were reliably different we have our definition. And, in the crudest sense we have a relationship even with only two points along the physical scale having been presented.

In defining a lower absolute threshold we first explore a physical stimulus scale in an area where we expect the subject to be able to perceive the stimulus part of the time and not perceive it at other times. Thus, we relate the scale to responding or not responding. Then, by arithmetical operations we determine a single value above which we expect the subject to respond more than 50 per cent of the time and below which less than 50 per cent of the time. This point is the threshold. Again, in actually presenting the stimuli we would use a particular psychophysical method and this would be a part of the elaborated definition.

The term *closure* has been used rather widely by some psychologists. Our literary conception of the meaning of the term is that it represents a tendency to see incomplete forms as being complete. Now, can this concept be given operational definition? A study by Bobbitt (4) shows that indeed it can. He presented triangles with

on behavior of a *zero* amount of a dimension is compared with the effect of some finite amount. If there is a reliable difference in behavior resulting from these two conditions, the procedures used to derive it define the phenomenon. The symbols, E/C, refer to experimental and control conditions; the experimental condition is the one having a finite amount of a given stimulus condition, the control condition, zero amount. The symbols have been used by Marx (8), but his interpretation of the concepts derived is somewhat different from my interpretation. I will return to this matter at a later point in this book. It is sufficient to say at this point that operational definition of phenomena by these operations has been used as long as control groups have been used, which is a good many years.

I have a number of issues about operational definitions in general which I wish to bring up in this section. While definitions of this type provide essentially the experimental design for research, there are many possibilities in the operations for a confounding or confusing with other concepts. Therefore, the operations and their consequences must be thought through very carefully. My illustrations will include both the use of physical stimulus dimensions and psychological stimulus dimensions, but I will make no breakdown of this difference. In all illustrations it will be noted that the operations identify the stimulus variable which is essential to define the phenomenon and in the crudest sense, establish a relationship.

1. One of the oldest phenomena in the psychology of learning is *retroactive inhibition*. This phenomenon can only be unambiguously defined and demonstrated in actual research by the E/C-type of definition. The basic operations may be outlined as follows:

	Task A?	Task B?	Retention of A?
CONTROL:	Yes	No	Yes
EXPERIMENTAL:	Yes	Yes	Yes

With this paradigm we need add only that retention under the control condition must be reliably greater than under the experimental condition. If this is true, the phenomenon is demonstrated and the procedures define it.

Two additional comments may be made at this point. The above

Once more I must introduce a word of caution, and ask you to think back to the previous chapter where we dealt with unitary and complex dimensions. If our scaling techniques allow us to infer a reliable dimension this dimension is given operational definition by the procedures. Furthermore, if this dimension is now manipulated as a stimulus dimension and behavior is reliably affected, the relationship is given operational definition, whether named or not. But, if the dimension scaled is complex, as it may well be, the investigator must guard against concluding that he has demonstrated a fundamental law based on a relatively unitary stimulus dimension. If the stimulus dimension is complex we must be aware that we might be able to break it down into more unitary dimensions. But, at the particular level at which an investigator works his operational definitions are sound; he simply should not be blind to the fact that further scaling operations may allow him to arrive at definitions of more unitary dimensions and thus indicate further research as to the relationship between these dimensions and behavior. If stimuli are reproducible, regardless of their complexity, and if reliable judgments of differences are made among these stimuli, then all criteria of operational definitions have been met. But let us not be myopic before the altar of operationism; let us recognize that it is an identifying device and that what one has identified is a matter for further research.

In this section on S-R identification I have discussed operational definitions of phenomena which are relationships between behavior and either a physical or psychological stimulus dimension. In these cases the "input" on the stimulus side was positive. That is, a certain measureable amount or quantity of the dimension was used. The phenomenon was defined if a minimum of at least two points along the dimension was sampled. At the same time, and even with only two points, a very crude relationship is established. Normally, however, in practice three or more points from along the dimension are used so that the relationship defined is more precisely given. The fact that at least two positive or finite amounts of the dimension are used contrasts this procedure with the final one which I will now discuss.

*S-R, E/C identification.* The operations under this heading differ from the other two S-R types in that in the simplest case the effect

such as *aggression* and *withdrawal*. It is also true that in a single experiment several different response measures may be taken, each, when noted in conjunction with the blocking operations, defining independent phenomena. This, of course, is quite legitimate. Indeed, it is necessary if the response measures are not highly correlated. But, the important matter to which I wish to call your attention is that we have a grouping of operational definitions under a superordinate definition. That is, frustration is defined in terms of general operations and merely specifies that there must be a difference in behavior resulting from the two conditions. If one wishes, then, as a number of investigators have, the behavior may be broken up and measured along several different scales so that each is in turn capable of independent definition.

Perhaps I should mention once more that provisional definitions are provisional; there is nothing in the definition which says that the operations will inevitably produce a difference in behavior, and hence complete the definition of the phenomenon. This is why the verbal definition is an "if-if-then" type. This is particularly relevant when considering frustration since there have been a number of investigations in which no differences appeared as a result of the operations. When this happens it is quite incorrect to report that frustration did not result in any measurable differences in behavior. It is correct to say that the operations used failed to demonstrate a phenomenon which, if it had been demonstrated, would have been called frustration.

3. A general definition of *transfer* is as follows:

	Task A?	Task B?
CONTROL:	No	Yes
EXPERIMENTAL:	Yes	Yes

Any reliable difference in performance on Task B completes the definition of transfer. Now actually, because of our state of knowledge in the area, this general definition would normally be broken down into two subdefinitions on the basis of direction of difference in performance on Task B. That is, we would say that if performance under the control condition on Task B is better than under the experimental, *negative transfer* is defined. If the experimental

type of definition is what I call a diagrammatic type of operational definition. In such definitions we have the outline of the experimental design needed to demonstrate the phenomenon. We can, of course, construct these definitions in the more common verbal form. These definitions become the "if-if-then" type of verbal definitions, and sometimes may get a little dangling. For retroactive inhibition we would say: "If under one condition only Task A is given, and if under another Task A is followed by Task B, and if the retention of A is better under the first condition than under the second, then we have defined retroactive inhibition." Because I think the diagrammatic definition is somewhat less burdensome and perhaps somewhat more easily grasped, I prefer it and will use it almost exclusively in subsequent discussion.

The second comment concerns specification of the direction of the difference between the experimental and control conditions. In some definitions (to follow) it is not necessary to specify the directions of the performance difference. Clearly, however, in the case of retroactive inhibition, the control condition must result in better retention than the experimental condition. Indeed, if the experimental condition resulted in better retention than the control we would have defined *retroactive facilitation*.

2. I would like next to discuss the definition of *frustration*, and certain additional issues which it raises. In diagrammatic form, frustration may be defined as follows:

CONTROL: Goal oriented: No blocking

EXPERIMENTAL: Goal oriented: Blocking

If, as a consequence of these procedural differences, there is a reliable difference in behavior, frustration is demonstrated, and the procedures define it. In this particular instance, further elaboration of what is meant by *goal oriented* and *blocking* will probably be necessary. That is, the investigator would need to specify what particular goal or task has been set for the subjects and what particular technique was used for blocking attempts to attain the goal. And, of course, the particular response or responses measured will be a part of the definition.

In studies of frustration, a number of different response measures have been used. Furthermore, some of these have been given names,

years as a reliable phenomenon simply because it had not been operationally distinguished from other already established phenomena. The simple reason for its not having been so differentiated was failure to use a control condition in carrying out the operations. An adequate definition of reminiscence requires the following procedures:

CONTROL: Learning: Retention Test 1—Retention Test 2

EXPERIMENTAL: Learning: Retention Test 1-----Retention Test 2

The series of dashed lines in the experimental condition indicates that the time interval between the first and second retention test is longer for this condition than for the control condition. Now, if the retention on Test 2 is greater for the experimental condition than for the control condition, reminiscence is demonstrated and the procedures define it.

In the earlier work on reminiscence no control group was used; rather, reminiscence was said to have been found if retention was better on the second test than on the first test for the experimental condition alone. However, because the first retention test might well provide an additional learning trial, Retention Test 2 might have shown higher retention than Test 1 simply because of this additional learning, and not because of any facilitating process which was presumed to take place during the interval between the two tests. Indeed, if we knew the amount of additional learning provided by the first retention test, there may have been actual forgetting in the experimental condition. Recent research (1) has shown that this is what happens. In other words, the operations had not clearly differentiated reminiscence from forgetting. I shall discuss later in somewhat more detail this problem of priority of concepts.

5. I now wish to turn to certain problems of definition which may be illustrated by the discussion of two phenomena, *conditioning* and *pseudoconditioning*. First, let me give a general definition of conditioning, using CS for *conditional stimulus* and US for *unconditional stimulus*.

Present CS-US series?    Test with CS?

CONTROL:

No

Yes

EXPERIMENTAL:

Yes

Yes



produces better performance than the control, *positive* transfer is defined.

I have said that one of the virtues of operationally defining concepts by the diagrammatic technique is that the basic experimental design is given by the definition. However, we should not be lulled into thinking that such definitions automatically protect us from errors in carrying out the operations. For example, in the definition of transfer it is quite apparent that if we use different groups for the two conditions they must not differ significantly in learning ability. Or, if we use the same subjects in both conditions, any potential differences in difficulty of the material must be balanced out. Let it be evident that an operational definition only makes it clear in general how you are finding what you are defining; it assumes experimental and statistical competence to carry out the operations so that no confounding of variables is present.

4. While it may seem fairly evident in the illustrations given thus far that the control condition is an essential part of the defining operations, this has not always been the case. I want to discuss two cases as illustrations of where failure to consider a control group as a part of the definition led to, or may have led to, the defining of phenomena which did not exist.

In some of the earlier work on the Rorschach, someone colored portions of some of the cards with a garish red. In presenting these colored cards it was believed that the responses to them differed appreciably, indeed dramatically, from the responses to black and white cards. This alleged difference was called "color shock." Now, it seems clear that to differentiate responses to colored cards from those to black and white cards, hence, define color shock, an appropriate control must be used. This could be done as follows:

CONTROL CARDS: Without color

EXPERIMENTAL CARDS: Same as control with color

Now, if any difference occurs in the quantified responses to the cards, an operational definition of color shock is given. Actually, the investigator might specify that the difference must be in a certain type or kind of response. But until such operations had been carried out there could be no acceptable definition of color shock.

In the field of retention, *reminiscence* masqueraded for several

If in the case of conditioning our definition must be so constituted that the performance observed on the test trials is broken up into as many unique "parts" as possible. One phenomenon which can be in a sense "subtracted" from total performance is *pseudoconditioning*. It may be independently defined as follows:

*Series of US presentations?    Test with CS?*

CONTROL:	No	Yes
EXPERIMENTAL:	Yes	Yes

If on the test trials more responses occur for the experimental operations than for the control, pseudoconditioning is demonstrated. Note that the only difference between these operations and the ones presented for conditioning is that in the present ones the initial series does not have the CS included. Any conception of conditioning (learning) includes the presentation of the CS during the training trials. If pseudoconditioning occurs, more is "getting into" the performance from which conditioning is inferred than is intended by the conception of the process defined. Indeed, we might want another control in which only the CS was presented during the initial series, and it is a fact that in the study of certain conditioned responses (e.g., eyeblink) such a control is used. In short, the purpose of the control groups is to isolate the performance change which can be attributed only to the pairing of the CS and US during the initial series. When, therefore, we speak of the E/C operations, we do not necessarily imply that the C indicates a single control condition; it may necessarily indicate several, depending on the level of analysis which has taken place around a given performance change.

Now, it may well be that phenomena which have been independently defined may be found to be a function of some of the same variables and may be included under the same basic explanatory principles. These matters, however, must be kept separate from the sheer matter of definition. The variables of which the phenomena are a function are a matter for further research; the inclusion of them in the same explanatory framework is a matter for the theoretician as he examines the research.

6. I have said that independent phenomena may be defined when

If the test performance under the experimental conditions is better than under the control, i.e., if more responses are made to CS, the procedures define the phenomenon. A very comparable set of operations would define *learning* in a general way except that we would omit the specificity of the stimuli of the initial series and indicate only that practice trials were given on a task. In very few modern researches on typical learning problems is the control group actually used. The reason appears to be that learning (or conditioning), being such a pervasive phenomenon, would be known to occur in the experimental group and not in the control. In most learning situations we would expect the control group to have a zero measure of performance on the test trials. For example, if the material were nonsense syllables it is highly unlikely that the control subjects would get any of the syllables correct when they had never before been exposed to them. In short, with some well-established phenomena, the control group is not needed to demonstrate the existence of the phenomenon because it has been demonstrated so many times in the past.

But now, turning back to conditioning as such, there is reason to believe that a control group becomes an essential part of the defining operations even though the phenomenon of conditioning has been demonstrated many, many times in the past.

I have insisted that science is a series of analytical steps. One of the purposes of analysis is to isolate unique phenomena. The problem is somewhat comparable to that of reducing stimulus dimensions to as unitary a level as possible. On the response side, likewise, we must isolate unitary or unique phenomena. Suppose that the phenomenon defined as conditioning is not unique in the sense that it is constituted of two or more isolable phenomena which can be given independent definition. Now, on a strict operational level, the definition of conditioning as I have given it would not be misunderstood; the operations are clear. But, as scientists engaged in an analytical enterprise, our definitions must keep pace with our analytical research. If, therefore, we can break a phenomenon down into more elementary ones by appropriate differentiating operations, we must do so. Or, to say this another way; we break the variance down into as many components as there are discriminable operations that affect it.

mobile so that no learning which would transfer could take place. Obviously, there would be objections from certain elements in our society toward the anesthetization of young children over a period of several weeks. Again, I must caution you against concluding that a phenomenon of maturation in human subjects does not exist; I am merely pointing out that this particular type of operation for defining maturation is unacceptable and insofar as it is unacceptable, maturation as allegedly defined does not have scientific status.

With this illustration in mind, let us try to arrive at some generalizations concerning the operational method of distinguishing among phenomena. It will be noted that in the discussion of maturation and transfer the problem of differentiating them must be done on the basis of stimulus manipulation differences, since the same response measure is used to define both. And this is exactly where the trouble lies; maturation and transfer were not differentiated because the stimulus manipulations were not differentiated; those used to define maturation are essentially the same ones used to define transfer. Under such circumstances we must observe a rule for priority of concepts based on our conceptualization of the types of processes involved. Priority goes to the concept which can be demonstrated in a situation in which the other, conceptually speaking, could not occur. For example, we can demonstrate marked positive transfer from one task to another in a 25-year-old man in five minutes. No current conception of neuromuscular maturation, that is, no usual way of thinking about the process, would predict this.

Let us consider another illustration of the problem. Zeller (15, 16) has done careful experimental work in an attempt to establish repression as a scientific concept. However, when he finished the two experiments, he realized that the performance changes which he measured might well be attributed to motivational changes of the subject. The motivational changes could well have been produced by the stimulus operations by which he intended to influence another process. Motivation as a concept has independent status; that is, the phenomenon can be defined in a situation where the literary conception of repression would not predict the appearance of repression. Zeller rightfully, although somewhat ruefully I believe, concluded that his operations were inadequate to establish the reliability of a new phenomenon.

clearly differentiating operations are used. I want to discuss this problem at some length, and will use initially as an illustration the problem connected with a definition of maturation. One definition which has been widely used is that resulting from identical-twin studies. One twin is used in a control "condition," the other in an experimental "condition," as follows:

*Practice on specific skill    Test on this skill?*  
(e.g., climbing)?

CONTROL:	No	Yes
EXPERIMENTAL:	Yes	Yes

If performance of the experimental twin has improved from the beginning of practice to the tests, and if there is *no* appreciable difference in the performance of the control and of the experimental twin on the test task, it has been said that maturation is demonstrated. The idea is that neuromuscular development sufficient to account for the behavioral change takes place without the specific practice. Unlike all of our previous definitions, the phenomenon rests on demonstrating no difference in performance on the test trials. As will be seen, my objection to the definition has nothing to do with the statistical impossibility of confirming the null hypothesis. My objection is that these operations do not allow definition of an independent phenomenon clearly distinguished from another well-established one. In fact, I would insist that insofar as the above type of operations have been used to define maturation, no such phenomenon exists in a scientific sense.

A well-established phenomenon in motor learning is that of transfer of skill. With this phenomenon in mind let us look at the control condition used to define maturation. The control group subjects are not kept immobile; as a matter of fact, they are allowed a great deal of activity. They may run, jump, crawl, turn somersaults, and so on, but they are not allowed to practice climbing. It is quite reasonable, therefore, that there could be heavy transfer from these other activities to climbing. If the control subjects, therefore, do as well as the experimental subjects on the test trials this might be attributed to transfer, and no new phenomenon need be defined. It becomes apparent that to give an acceptable definition of maturation by these types of operations the control group must be kept completely im-

the practices of operationism. I will build this summary around these three benefits.

One of the benefits stated was that operationism facilitates communication. I can do no more than indicate that I think this is self-evident. How can the sheer empirical meaning of a concept be misunderstood when it is operationally defined? I can only insist that the variance in the transmission of meaning by operational definition is far, far less than that for literary definitions. I think that any problems which have arisen over this matter have come about because some have expected definitions to do more than define. One must not look for hidden meanings in operational definitions; there are none. The entire empirical truth status of a concept is given by the defining operations, and that is all that a definition is supposed to give. One who reads an operational definition in the spirit in which it has been formulated has no problem in understanding what is meant by it.

A second benefit which I have listed as resulting from the use of operational definitions is that it makes better scientists out of us. When the meaning of a concept is evaluated by the criterion of operationism it forces us to sharpen our research designs. Unless a phenomenon, denoted by a set of operations, can be distinguished operationally from already existing concepts, we do not admit it to the empirical base of the science. A secondary effect of this insistence on distinguishable operations is that it keeps the number of concepts limited, at least for certain types of operational definitions. This is the next problem to which I will turn.

The third benefit of operationism, as I see it, is that the number of concepts admitted to a science is less under operationism than under, say, literary types of definitions. At least one can state criteria by which concepts may or may not be admitted to a science under operationism, whereas this is not true with literary definitions. However, I would like to summarize the issues bearing on this matter at some length. To do so, I will mention separately the problems as they relate to the six types of operational definitions discussed earlier.

In my opinion the area where the pyramiding of operational definitions might get out of hand is in simple response-defined concepts. The problem stems from the fact that in this area there is no

We may then state a general principle of precedent or priority. If the same response measure is used in the defining operations of two phenomena, and if the stimulus manipulations cannot be clearly differentiated, the phenomenon which can be demonstrated (hence defined) in a situation where by its literary conception the other would not occur, the first phenomenon takes precedent. I do not think this rule of precedent will solve all definitional problems where stimulus manipulations conflict, but I think it will handle most of them.

All of this discussion is concerned with differentiating phenomena when the same response measure is or was to be used; hence, differentiation must be based on the stimulus manipulations. This does not close the problem for these E/C type operations. For, there are instances in which the same stimulus manipulations are used for two or more phenomena and in which case the differentiation must come in the response measurements. The rule here is fairly straightforward. If the same stimulus manipulations are used, phenomena are differentiated by different response measures if, and only if, those response measures are poorly correlated. The only ambiguity in this principle is that we have no set value of the correlation coefficient which can be used as a clean-cut criterion as to when and when not responses are said to be poorly correlated.

We have previously given a definition of frustration, and saw how this general definition can subsume subphenomena, such as aggression or withdrawal. In these instances the stimulus manipulations used to produce aggression and withdrawal are the same; they are differentiated on the basis of response measures. Obviously, if these response measures correlate highly there is no basis for distinguishing two phenomena. But, since they do not correlate highly, scientific analysis is aided by insisting upon definition of independent phenomena.

#### A SUMMING UP

In this final section I would like to bring together certain of the scattered comments in the chapter and add a few statements of appraisal. Early in the chapter I indicated the three major benefits which I believed accrued to psychologists accepting wholeheartedly

matter; rather, it is a research matter. And I would say, also, that I know of no other technique of defining concepts other than the operational technique which can insist upon a criterion for the admission or nonadmission of concepts.

I will turn next to definition by S-R identification. It will be remembered that concepts defined in this category always require the designation of a particular stimulus variable responsible for the phenomenon being defined. Pitch is defined only when variations in cycles per second of the sound wave are "put into" the organism and reliable judgments are obtained. Such variations in cycles per second can be produced in a number of ways, from the use of simple tuning forks to complex oscillators. If the precise relationship between the phenomenal change in sound and cycles per second changes with the particular techniques of producing the sound or of measuring the subject's responses, we do not need separate concepts for each minor deviation. It poses an interesting research problem as to why the methods do not produce the same relationship, but this is not a definitional issue. In other words, we have a basic criterion for including what could be many, many variations in a class when they meet this criterion for inclusion (11). In this case, the criterion for inclusion is that there be variations in cycles per second and reliable differences in judgment of the sound.

Operational definitions based on S-R relationships and a psychological scale depend for their adequacy on the operations used to define the psychological scale. Hence, the correlational criterion must be applied to assess the number of concepts to be allowed.

This brings us finally to S-R, E/C identification. Concepts resulting from these operations rest on the demonstration of a difference in performance as a result of a finite amount of a specified stimulus condition being present and performance when none of the condition is present. Thus, as in the case of identification with a physical scale, the definition rests on the specification of a variable necessary for the phenomenon to occur. Retroactive inhibition, for example, is defined in terms of a performance difference in retention when interpolated learning is and is not introduced. Any kind or amount of material so interpolated would fit the class of operations defining the concept. It is irrelevant to the general definition how fast the material is pre-



defining criterion as such which tells us whether or not a new concept should be admitted. The only criterion which can be used is a research criterion. This criterion, stated simply, is that no new concepts should be admitted if the phenomenon is already measured by what might *appear* to be a somewhat different set of operations. You will remember that test-construction procedures illustrate well the simple response-defined concepts. Almost an infinite number of tests could conceivably be constructed and if one stopped at that point, each would define a new concept. We have only one protection against such an eventuality, namely, insisting that correlations be determined among tests and not admitting a new concept if the characteristic involved has already been defined. That is, if the correlation is high between one test and another, the same concept should be used for both. The fact that we have a number of reliable tests all called intelligence tests shows that this is working out to a certain extent.

While I would insist that this correlation criterion provides the only adequate protection against a wild proliferation of concepts in this area, I must quickly add that it is not simple to work out in practice. Suppose we have a test which defines the concept of *clerical aptitude*. Then suppose another investigator constructs a test to define *finger dexterity*. It might never occur to this second investigator that his test correlates very highly with the one defined as measuring clerical aptitude, although this might actually be the case. In short, there are hundreds of tests which have been constructed to define certain concepts and it would be next to impossible for an investigator to correlate his new test with all of these. We may, therefore, expect that tests which actually measure the same thing will have different names applied to them.

It is in the handling of this whole problem that I have mentioned that factor analysis makes a very strong contribution; factor analysis limits the number of characteristics of behavior which must be given independent definition. So the problem is not hopeless; it is only a huge research task. The only other hope is for occupational fatigue of test constructors. I must repeat again that we must have independent response-defined concepts if the response measures for two or more tasks or tests or situations do not correlate. Restricting the number of concepts defined by response measures is not an arbitrary

"introversion as a function of number of siblings." Of course, the phenomena which become central may vary considerably from scientist to scientist.

2. Although I have tried to guard against it, it is still possible that the temper of this chapter may have left the impression that our operationally defined phenomena all have about the same status as far as generality is concerned. This is obviously not the case as indicated by the discussion of class operations. Let it be clear that we do have widely different levels of generality in our operations. For example, we may have a very general definition of forgetting, but under this we have a set of operations defining retroactive inhibition, and under this a set defining the influence of similarity on retroactive inhibition and under this a set defining a particular kind of similarity as it might be varied in a given experiment. We have classes within classes within classes.

3. Operational definitions as I have discussed them in this chapter are used to define behavioral phenomena. This is a somewhat more restricted use of operational definitions than I believe is common in psychology. However, I do not want to discuss this matter now since in terms of my thinking it comes more logically in Chapter 7 where the status of varying kinds of concepts in use in psychology will be evaluated.

## REFERENCES

1. AMMONS, H., & IRION, A. L. A note on the Ballard reminiscence phenomenon. *J. exp. Psychol.*, 1954, 48, 184-186.
2. BERGMANN, G., & SPENCE, K. W. Operationism and theory in psychology. *Psychol. Rev.*, 1941, 48, 1-14.
3. BILLS, A. G. Changing views of psychology as science. *Psychol. Rev.*, 1938, 45, 377-394.
4. BOBBITT, J. M. An experimental study of the phenomenon of closure as a threshold function. *J. exp. Psychol.*, 1942, 30, 273-294.
5. BOUSFIELD, W. A. The relationship between mood and the production of affectively toned associates. *J. gen. Psychol.*, 1950, 42, 67-85.
6. GRAHAM, E. E. Wechsler-Bellevue and WISC scattergrams of unsuccessful readers. *J. consult. Psychol.*, 1952, 16, 268-271.
7. LEVINSON, D. J. An approach to the theory and measurement of ethnocentric ideology. *J. Psychol.*, 1949, 28, 19-39.

sented, how high the degree of learning is, and so on, if the specified performance difference is found. All other factors mentioned may indeed be variables affecting the amount of retroactive inhibition but they are static variables as far as the defining operations are concerned.

Finally, I am compelled to make additional comments partly by way of preparation for later chapters, partly by way of keeping some pesky terminology problems from getting out of hand, and partly to correct any erroneous impressions that empirical concepts are all nicely organized. What I have to say is concerned largely with S-R types of phenomena.

1. I have said, and will continue to say in later chapters, that at the empirical level our research problem is to determine behavioral phenomena and variables which influence them. I have also said that an operational definition specifies the necessary operations needed to demonstrate the phenomenon. Now this word "phenomenon" occurs hundreds of times in this book; I don't like the word but I haven't a good substitute. I think it is clear that when I use the word I mean a reliable behavior event or change; it is an event whose recurrence can be discriminated as such. When I say "phenomena and variables affecting these phenomena" I am stating my own bias for organizing research findings but I think I am also reflecting the organizational structure of much research in psychology. Research workers *do* center their work around one or two phenomena as they go about determining the influence of variables on these phenomena. But, to speak in this way is no more than that; that is, it is just a manner of speaking. For any reliable behavioral event or change is a phenomenon so that these "variables which influence phenomena" do themselves define phenomena. So perhaps I should say "core or central phenomena and associated phenomena" instead of "phenomena and variables which influence them." In any event, I think it is fair to say that many psychologists do think in terms of core phenomena and that other phenomena revolve around them; these others I will continue to refer to as "variables which influence them." This centrality is shown by titles to research papers, to section headings in books, and so on. Thus we read: "extinction as a function of amount of work;" "pitch as a function of intensity;" "intelligence as a function of race;" "learning as a function of meaningfulness;"

# *Research Design: I*

## INTRODUCTION

In Chapter 2, I discussed the components of the research situation in psychology, and made some preliminary remarks about the research implications of these components. In Chapter 3 I was rather doggedly dogmatic in my presentation of operational definitions. Yet, following the manner in which I presented these definitions, it is quite natural to go now into the matter of research designs. As I have indicated, the operations required for demonstration, hence definition, of a behavioral phenomenon are contingent upon certain principles of procedure which, if overlooked or misused, deny us the privilege of definition. I, therefore, feel that the base of our science (reliable phenomena) exists on a firm foundation only insofar as our research procedures are sound.

I must first give you a general idea of the approach I shall take in discussing problems of experimental design. The material will not be presented simply as an exposition of different research techniques. It is my experience that there is a certain flatness to such an approach; it does not quite hit the mark that I think it is necessary to hit. I gave an opinion earlier that we are far too uncritical of our research efforts. Many published research reports should have been thrown in the wastebasket rather than have been mailed to an editor, for the editor is sometimes so harried by the tide of papers seeking readers through his journal that he may find time for little more than a cursory examination of their soundness. I must repeat that the blame, if we must cast blame, rests unequivocally on our graduate training programs.

If you want basic principles for designing research problems they can be given rather simply; rather "it" can be given rather easily.

8. MARX, M. H. Intervening variable or hypothetical construct? *Psychol. Rev.*, 1951, 58, 235-247.
9. MONTGOMERY, K. C. The relation between exploratory behavior and spontaneous alternation in the white rat. *J. comp. physiol. Psychol.*, 1951, 44, 582-589.
10. SPIKER, C. C., & McCANDLESS, B. R. The concept of intelligence and the philosophy of science. *Psychol. Rev.*, 1954, 61, 255-266.
11. STEVENS, S. S. Psychology and the science of science. *Psychol. Bull.*, 1939, 36, 221-263.
12. TAYLOR, J. A. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, 48, 285-290.
13. TOLMAN, E. C. The nature and functioning of wants. *Psychol. Rev.*, 1949, 56, 357-369.
14. WATERS, R. H., & PENNINGTON, L. A. Operationism in psychology. *Psychol. Rev.*, 1938, 45, 414-423.
15. ZELLER, A. F. An experimental analogue of repression: II. The effect of individual failure and success on memory measured by relearning. *J. exp. Psychol.*, 1950, 40, 411-422.
16. ZELLER, A. F. An experimental analogue of repression: III. The effect of induced failure and success on memory measured by recall. *J. exp. Psychol.*, 1951, 42, 32-38.

publication by the editor (who does so without bringing the associate editor or editors into the picture by name). If the author hasn't already guessed that I was involved in the rejection of his manuscript, he may now find this out. [The rejection of my own manuscripts has a sordid aftermath: (a) one day of depression; (b) one day of utter contempt for the editor and his accomplices; (c) one day of decrying the conspiracy against letting Truth be published; (d) one day of fretful ideas about changing my profession; (e) one day of re-evaluating the manuscript in view of the editors comments followed by the conclusion that I was lucky it wasn't accepted!]

Finally, since I have seen many research proposals *de novo* from students, I shall mention them when they add something new to our thinking.

Much of the research about which the discussions will center is research of an experimental nature, i.e., where some variable is manipulated. It is in this type of research that it is easiest to go astray. I believe that if we understand the problems associated with design of experiments we can readily transfer to other forms of research, e.g., statistical control of variables, naturalistic observation, and so on. However, research involving simple and complex response-defined phenomena will not be covered directly.

Some research workers in psychology assert that design of experiments and statistical evaluation are inextricably related. I believe that they can be productively separated. I have seen a sufficient number of reports which are statistically sophisticated and behaviorally naive to lead to this conclusion. I will not, therefore, consider statistical problems except in very minor ways. I am concerned with the thinking that goes into experimental designs; such thinking may at times overlap statistical thinking but it is propaedeutic to it. If, however, the presentation is incomplete because of the failure to handle the statistical problems, then so it will have to be. In many cases a design may be changed to allow better or more complete statistical evaluation. These are the details which I will leave to the statistician. So also will I leave many other details of experimental procedure to the experts in the particular area of research in which one may be interested. I cannot tell you how to wire a six-channel recorder nor can I tell you how to handle schizophrenic patients in an experimental situation. There are many, many such details,

For there is only one basic principle, namely, design the experiment so that the effects of the independent variables can be evaluated unambiguously. The difficulty lies in implementing the principle in any specific research situation. I do not think there is any easy solution to this learning problem. I do not think, furthermore, that we can simply tell research workers to "watch out" for infraction of the rule and expect positive results. In order to learn we must practice and practice and practice. We must learn what not to do as well as what to do. In the present work I can no more than get one started on this learning process; it is up to each of us to continue the practice.

Essentially what I will do is give you illustrations of research errors. Then, I will note how each error may be corrected. Some of the illustrations will be from published reports; I use these reports because they are real and can be evaluated by all who choose to do so. As I said earlier, I think it is somewhat unfortunate that these should be found in print; it is at least unfortunate for contemporary psychology if it is going to be evaluated for scientific rigor. If there is a positive side it is that such published reports may in the long run make better scientists out of all of us. If errors are to serve as a learning device, they cannot be hid. I doubt whether there is a single psychologist, actively engaged in research work, who has not at some time at least planned an experiment which did not meet critical standards. Most have probably executed such an experiment and some of us have mailed them to an editor. A critical colleague or a friendly editor (there are some) may have saved the research from being published. So, when I evaluate critically a piece of published research I do so without malevolence and with full awareness that many of us are in the fortunate position of having uninhibited colleagues and students who take fiendish delight in pointing out our blunders, usually before they become public.

I have also had available another source of research reports which has provided me with a number of illustrations of errors on which to focus when designing research. As a consulting editor for the *Journal of Experimental Psychology* for six years I have had the privilege of reviewing scores of manuscripts. I cannot, of course, give references to these works. But I shall reconstruct them from my notes as faithfully as I can. Some of these have been rejected for

and one which carefully analyzes what each type of research can provide (7).

*Organization scheme.* In the several years that I have been teaching research design to first-year graduate students no single problem has caused me more persistent anguish than that of trying to organize research errors into a meaningful pattern. I would like (as I think anyone would like) to have a scheme of presentation which is logical and at the same time exhaustive of these research errors. It might be useful too to have a checklist of known errors so that in designing an experiment we could go through and see if we have successfully avoided such errors. This assumes that we are perceptive enough to tell whether or not an error is occurring; knowing what errors can occur is not necessarily identical with perceiving when they do occur. So, the checklist idea does not solve the problem; it may help by calling attention to possible errors which we might have otherwise overlooked but it will not help us in determining whether that error is or is not present in a given investigation. What I am leading up to is a statement that I recognize my presentation is neither completely logical nor exhaustive; it is the best I have been able to formulate and unless a flash of insight strikes me before the final page proof leaves my desk we are stuck with its deficiencies. Nevertheless, in one way or another I think I shall hit upon most kinds of errors which do occur in psychological research. Let me then turn to the thinking which lies behind the organization.

What do I mean by errors in research? To determine whether or not there is an error in an investigation requires a comparison of what the investigator does and what he concludes was found as a consequence of this doing. Experimental research involves the manipulation of a variable in some fashion; the intent is to discover if this manipulation is related to behavioral changes. What the investigator said or implied he has found in light of what he did are the two critical focal points in determining whether a research error does or does not exist. (If we are evaluating proposed research the problem is no different; we look at what is to be done and what conclusions are intended, given various alternative findings.) If there is a discrepancy between what *is* concluded and what *can* be concluded in the light of what was done, I assert there is a research error. In terms of seriousness, these discrepancies vary considerably. Some



important indeed, which only the active worker in the particular area can give with the degree of expertness required. I am not aiming at this level of research design or procedure.

I suppose if I get clearly impatient on the matter of research design it is in the case where before an experiment is done it is recognized that the answer to the question cannot be obtained or a test of the hypothesis accomplished but the research is undertaken anyhow after uttering the soporific phrase "it is the best that can be done under the circumstances." I would insist that the research should not be done under this aegis. Why do it if it cannot possibly solve a problem or answer a question? It is not unusual to carry out an experimental procedure and then discover that it does not accomplish what we thought it would, but to do the experiment when we know it won't accomplish what we want to accomplish is a clear expression of research stupidity or an unwarranted faith in the virtues of science.

It will be apparent that my presentation deals almost exclusively with the classical nomothetic approach in which groups of organisms are used. I have been unable to sense the revolution taking place in psychological research which others have seen. This alleged revolution deals with the study of the individual and is sometimes called idiographic psychology. The idea is that we should discover the laws holding for the individual, not the laws holding for the group. I cannot see how anyone can object to the study of a single individual; he may be studied intensively in the sense that many different relationships are determined for him or intensively with respect to only one particular phenomenon. The latter plan of study has long been used in research on sensory processes. Ebbinghaus used it for his studies of memory and Skinner uses it for the study of a rat. I do not see that there is a systematic issue involved here unless those who champion idiographic analysis are trying to say that there is no commonality of laws from one organism to the next, in which case we will have as many sets of laws as we have people. I would insist that the laws and relationships about which we already know would deny this stand. So, where is the issue other than the ever-present one in all kinds of research concerning the generalizability of results. Nevertheless, since I may be missing something I would refer you to two "pro" idiograph papers (2, 13), one "anti" (14),

class the results may be confounded by (a) other identifiable variables within the same class, or by (b) variables from the other two classes. In terms of research errors actually made, however, certain confoundings are more probable than others. I say that with a conviction unwarranted by any information I have available, for I have neither taken a random sample nor made a systematic count of frequencies of various types. Yet, I think it is true and the length of my discussions for various types of confoundings will reflect this belief.

May be Confounded by . . . Variable

		Environmental	Task	Subject
When Manipulating . . . Variable	Environmental	1 X	2	3 X
	Task	4	5 X	6 X
	Subject	7	8	9 X

In the accompanying table I have expressed the fact that when a variable within a specified class is being manipulated the results may be potentially confounded by other variables within the class or by variables from other classes. The cells marked with X represent confoundings which I believe occur with greatest frequency. With this table in mind, I will indicate the order of topics to be discussed.

discrepancies are relatively unimportant in the sense that verbal conclusions are slightly askew but can be easily corrected and we retain the research as a sound contribution to the literature. At the other extreme are discrepancies which are lethal and there is no way that a scientifically meaningful conclusion can be reached from the procedures used. These cases are best exemplified by blatant confounding of stimulus variables from different classes (environmental, task, subject) so that behavior changes measured cannot be said to be the result even of variables within a given class. In between these two extremes are discrepancies which, to be charitable, reflect a nonanalytical conclusion in an area where development of knowledge is far enough advanced to disallow such conclusions. Such conclusions are disallowed because there is a confounding of the manipulated variable by another variable in the same class. What happens is that as the investigator manipulates one variable in a class, another identifiable factor in the same class also changes so that the phenomenon produced cannot be said to be due to one particular component; it can only be said to be due to the one or the other or the combined influence. A design error occurs if identifiable components of the manipulations are not isolated when it is quite apparent that they must be if analytical conclusions are to be drawn. Note that the confounding is between variables within the class in contrast to the confounding of variables from different classes as discussed above. With some justification, confounding of variables within a class is considered a less serious error than confounding of variables among classes. However, we shall have plenty of opportunity to compare the two and you may arrive at a different assessment of relative seriousness.

It is my belief that the major errors in psychological research lie in these two kinds of confoundings and my effort is directed toward extensive discussion of how such confoundings have occurred and how we go about trying to avoid them. The organization scheme for this presentation can now be outlined. It will be remembered that I identified (Chapter 2) four types of variables which may be manipulated, namely, *environmental*, *task*, *instructional*, and *subject* variables. For this discussion I will omit the instructional variable as an independent class and include it under the environmental variables. I have said above that when we manipulate a variable within a

## "RANDOM" GROUPS

On statistical grounds we have no better way of forming groups (which we wish to be statistically equivalent) than by assigning the individual subjects to the groups on a random basis. This may be in terms of sampling from a population or it may be (and this is more usual) assigning a limited supply of subjects to different groups without particular reference to the population of which the subjects might be a sample.

Statisticians have paid considerable attention to the matter of deriving random groups, with indications of what may be considered random and what may not be considered random, so I will not dwell on this matter. It is sufficient to say that we still have investigators who pay little attention to this issue even though it is an extremely critical one. The logic behind random groups is simple but powerful; if subjects are assigned at random, differences in groups on any subject variable are highly improbable, the probability being expressed by sampling error theory. If a bias enters the procedure of assigning subjects to groups, then differences in skills may exist between the groups which are relevant to the performance on which the groups are to be measured during or following differential treatment.

The policy of some journal editors has grown to be one in which they essentially refuse to accept a statement by an investigator that "subjects were assigned to the various conditions on a random basis." Or perhaps the better statement would be that the editorial policy requires the investigator to state exactly how his subjects were assigned to groups and then the editor decides whether the method could or could not introduce a bias. This policy has my full support. For in spite of the statistical elegance of random groups it is a fact that achieving true randomness in the experimental situation is not always easy. If there is a question of doubt concerning whether the method of assigning subjects results in random groups, it is the obligation of the investigator to convince the reader, by use of supporting data or logical considerations, that his results have a low probability of being biased by the method of forming groups. We can ask no more, for strict randomness gives us no more, but we can ask this much in the interests of rigor in our science.

One of the most persistent and deadly errors which occurs in research is that when manipulating an environmental or task variable, subject variables also change. This is the problem of getting and maintaining equivalence of groups of subjects and will be the first topic; this refers to cells 3 and 6 in the table. Cell 9 is a special case; it refers to research in which a subject variable is manipulated but in which the results are confounded by the simultaneous variation of other subject variables. A consideration of cells 3, 6, and 9 will complete this chapter. In the next chapter I will consider in order cells 1, 2, 4, 5, 7, and 8.

While I consider the various stimulus confoundings to be the major source of error in psychological research there are a number of other problems involved in the research process which, when not adequately handled, may also be said to constitute errors. These will be considered following the extended material on stimulus confounding.

#### THE PROBLEM OF EQUIVALENT GROUPS WHEN MANIPULATING ENVIRONMENTAL AND TASK VARIABLES

In the usual situation where environmental or task variables are manipulated, two or more groups are treated differently with the aim of arriving at a conclusion concerning the effects of the different treatment. To reach this conclusion, the skill or abilities of the groups *per se*, and the experiences these groups have, should not differ except for that inserted by the investigator as the experimental treatment. A major source of error in carrying out research is failure to appreciate the importance of this simple principle. Obviously, causes of differences in ability levels and interaction of differences in ability level with differential experimental treatment is a legitimate area of study, but I am concerned now with the case where the logic of the research rests on equivalent groups and the investigator is interested only in the effects of manipulating an environmental or task variable. As with most of the errors we are studying, we have obvious infractions and others that are quite subtle.

As I have indicated several times, experiments should not be open to the suspicion that groups may not have been equal before experimental treatment. But we will find in the published literature reports in which the results are such that we are almost forced to conclude that the groups must not have been equivalent. Of course we expect that random assignment will result in nonequivalent groups with a low frequency as indicated by sampling error theory. But, disallowing this infrequent occurrence (and we simply have to live with these) there are reports which, looking at the findings, lead one to conclude that random sampling error could not possibly be responsible for the findings. Let me give an illustration.

A study was set up to determine the influence of temporal point of interpolation on retroactive inhibition (12). For all conditions 16 days elapsed between original learning and relearning. There were nine conditions differing only in the point at which the interpolated list was inserted between original learning and relearning, the extremes being just after original learning and just before relearning. The original and interpolated lists were presented for a constant number of trials, the same number to all groups. A total of 63 subjects is said to be assigned randomly to the conditions so that seven subjects occur in each condition. The resulting curve of retention as a function of point of interpolation is quite striking, with three distinct peaks and with evidence presented that some of the differences between extreme points on the curve are highly significant statistically.

As a psychologist who has done work in the area of retention, I am particularly interested in this study. I seriously doubt that the behavioral law resulting from the manipulated variable is as complex as indicated by these results. However, the point I wish to stress is that such a doubt could have been easily removed by the investigator. It is well known that retroactive inhibition is a function of degree of learning of interpolated material and degree of learning of original material. Suppose that the small groups differed in amount learned under the constant number of trials. Certain of the results may be accounted for by this rather than by the manipulated condition. To allay such doubts all the investigators had to do was present evidence on the level of learning attained by all groups on the original list. Such data were obtained as a normal consequence of the

I have seen experimental reports which, simplified, had methods of forming groups about as follows. One entire class of elementary psychology students, meeting each day at 8:30 A.M. for class, is given one experimental treatment. Another entire class, taking the same course but meeting at 1:30 P.M., is given a different treatment. Differences in behavior are then said to be a function of differences in treatment. I am sure that you can think of a number of reasons why we would not expect these two groups to be the same as if we had thrown all the students in both the classes together and assigned them to the two treatments on a random basis. Depending on the nature of the performance being measured we may even point out specific reasons why we would expect one group to be superior to the other. I do not think we can accept such research unless the investigator shows that the two groups did not differ on a performance that is relevant to (correlated with) the performance measured during or following differential treatment. Thus, in the above illustration, the investigator might use a pretest to show that the groups did not differ on, say, attitude toward authority before he introduced differential treatment designed to change attitudes toward authority. This is what I mean by the investigator supplying supporting evidence of equivalence of groups when his method of deriving his groups is suspect as far as the random-groups logic is concerned.

I have also seen studies in which students in one school, say Commerce, are used for one condition and students in Liberal Arts for another without adequate supporting evidence that the groups did not differ appreciably on relevant variables. I have also seen research reports in which subjects were obtained on a semivolunteer basis and in which the first 50 subjects to volunteer were placed in one condition, the second 50 in the next, and so on. We cannot compromise this issue; the method of assigning subjects must be assuredly random or the investigator must present supporting evidence that whether random or not the effect was the same—the groups did not differ significantly on variables relevant to the skills required for the experimental task. The issue is such a simple one that I sometimes think we overlook it in our concern with the complexities of the details of the treatments given the various groups and in our concern with statistical analyses of the data obtained.

in which we used a practice task and an experimental task with the idea of matching groups on the practice task. Even though I believe that anyone would agree that these tasks had high apparent commonality the correlation between the two was found to be much too low to justify use of the practice task as a matching task.\* Thus, even though the groups did not differ appreciably on the practice task the small differences which we observed on the experimental task may be, at least in part, a function of unequal ability and not the experimental variable.

*The ex-post-facto "experiments."* These experiments are based on the same reasoning as are experiments involving active manipulation of a variable. The idea is to search records which have been kept, segregate out two or more groups which were at one time equivalent but which subsequently were exposed to different "conditions" or "treatments" through the natural order of events. These groups are now measured and if differences appear they are attributed to the different treatments which had occurred. This is exactly what we intend to do if we actively manipulate a variable following our assigning of groups at random or following the matching of groups on a relevant variable. So, we have no new ideas involved. But, the fact is it is virtually impossible to carry out such research which meets standards for drawing cause-and-effect relationships such as we arrive at when we actively manipulate the variable. We need to examine the problems in some detail. I bring it up at this point because certain investigators have apparently made the assumption that matching of subjects extricates them from the hopeless situation which usually obtains for such research.

Let us start with an ideal situation. Suppose that in 1940 we formed two random groups from a defined population. Then suppose that a very unplausible event occurred in 1942. All of those subjects in one random group were drafted into the Navy and all subjects in the other group were drafted into the Army. In our fantasy let us also assume that in 1946 all members of both groups were still alive and discharged and we asked the simple question. "Is there a difference in frequency with which the two groups were enrolled in colleges and universities?" If there is, we might well attribute it to

\* *Mea Culpa.* In this study I am embarrassed to say that we did not specify how subjects were assigned to groups—only that they were assigned at random.



operations and should have been presented as a matter of course to support the implication that the subjects were assigned at random.

If I may summarize, I think there are two principles we should follow when reporting research in which random assignment of subjects has been used:

1. State exactly the procedure used to assign subjects to the groups, e.g., random numbers, alternation, etc.
2. If possible, that is, if allowed by the experimental design, give data which support the expectation of equivalence of groups when random assignment is made.

### MATCHED GROUPS

If an investigator does not have a way by which subjects can be assigned at random, or, even if he does, he may prefer to use a matching procedure whereby the groups are equated on a relevant skill before introducing the experimental treatments. I have discussed details of such matching procedures elsewhere (15) and will not repeat them in full here. I wish only to hit some critical points which still are sometimes overlooked in current investigations.

In the first place, matching must be on a relevant task, skill, performance, or whatever is involved in the research. This means simply that the matching scores must be related—correlated—with the performance that is measured during or following the introduction of the independent variable. Just how high this relationship must be is again a question which cannot be given a general answer for the same reason as discussed in Chapter 2 when response reliability was considered. Some investigators still have a tendency to *assume* that tasks are correlated without a statistical demonstration of the correlation. If the matching and experimental tasks are not correlated we are forced to resort to an assumption of randomness which will usually be a highly questionable assumption since subjects are lost during the matching procedure. Matching subjects on intelligence before introducing them to differential treatments in a rote-learning task is questionable but it has occurred rather frequently without any apparent concern that intelligence test scores and rote-learning skill may be poorly related.

We have performed some studies on concept learning (e.g., 11)

before they actually started scouting. One could match groups on 1,000 other variables and still the fact cannot be gainsaid that they still differ on a relevant variable or variables which caused them to spend different lengths of time in Boy Scouts. As a matter of fact, in this particular case, the differences that were significant after 4 years (e.g., number of different organizations to which they belonged in 1938) appear quite reasonably to reflect the same fundamental difference which caused the boys to spend a different length of time in Boy Scouts in the first place. At best, then, all we can say is that length of time spent in Boy Scouts is correlated with subsequent community activity; we can predict one from the other. But, to imply any causal relationship between length of time spent in Boy Scouts and subsequent behavior is quite unjustified by the data.

So what are we to do when we want answers to such questions? As I have indicated earlier I have no patience with the sop that "it is the best we can do." If we can't do research better than this let us not do it. We must realize that there are important problems of behavior which cannot be attacked because of social mores. For example, if we wanted to get an answer to the Boy Scout problem we ideally would draw two or more random groups from the same population then force one group to spend so many years in Boy Scouts, another group a different number, another group not being allowed to join (control) and so on. Then we can measure differences after a given period of time and if we do not have differential mortality, we could ascribe any differences found to differences in length of time spent in Boy Scouts. I suspect such a procedure as outlined would meet some resistance in our society. Let us accept the fact that there are some problems for which we cannot find answers because the very nature of the problem prohibits us, not from adequately designing the research but from carrying it out. Although I am sure that I am not acquainted with all *ex-post-facto* studies which have been done, I do not know of a single one which meets acceptable research standards and from which cause-effect statements can be made. I shall return later to problems which arise in connection with the manipulation of subject variables which are similar to the problems which arise in these *ex-post-facto* studies. Let it be clear here that matching doesn't solve the problems of the *ex-post-facto* research.

differing events associated with Army *versus* Navy experiences. And I think that if such a difference were found we would be justified in reaching this cause-effect conclusion although we would realize that the particular nature of the differences in experiences could not be specified just from the data we have. This hypothetical experiment, in short, would meet our standards of design. But now let us see what is actually done in these *ex-post-facto* experiments. I will use as illustration an experiment concerning the length of time spent in Boy Scouts as it related to community adjustment (4).

It is a perfectly legitimate scientific question and an important social question to ask whether or not being in Boy Scouts influences later community adjustment or community participation as an adult. In this particular research, done in 1938, two groups of boys were separated based on the number of years spent in Boy Scouts through 1934, at which point all boys had terminated their association with the Scouts. One group had spent an average of 4 years, another 1.4 years in Scouts at time of termination. In 1938 these boys were measured on several factors related to community adjustment, community participation, and so on. It was clearly the intent of the research to establish a causal relationship between length of time in Boy Scouts and community adjustment. Obviously, not all boys were available in 1938 so the investigator resorted to a matching procedure on the boys who were available. (The problem of generalization of research findings is a matter we shall consider later.) The groups were matched on several factors, and even though the investigator did not report the relevance of the matching variables to community adjustment, this is not the principal point I wish to make. Are we to assume that the fact that one group spent 4 years and the other 1.4 years in Boy Scouts is due to sheer chance? I think we could agree that this would be highly improbable. The groups must have differed on one or more variables which are responsible for one group's being in Boy Scouts only 1.4 years in one case and 4 years in the other. If, then, we measure these boys 4 years later, we are probably measuring the continued influence of these factors and perhaps not at all any influence of different lengths of time spent in Boy Scouts. In short, we do not know that, and we do doubt that, the two groups of boys were random samples from the same population

probably accept the conclusion that meaningfulness was not a significant variable. But, if we find that the number of subjects unable to learn the list of low meaningfulness was greater than the number unable to learn the list of high meaningfulness we would insist that the experiment was not an adequate test of the influence of the variable. What probably happened is that we lost many more slow-learning subjects from one group than from the other so that our determination of differences in learning as a function of meaningfulness was confounded by differences in the learning abilities of the groups as they completed the experiment. Even if significant differences are obtained in the expected direction, if loss of subjects is different for the two groups we must recognize that the magnitude of the difference obtained is probably considerably underestimated.

This selection of subjects inevitably plagues the investigator when manipulating a task variable that does influence performance. The problem can usually be avoided by using a constant period of work or practice for the task rather than a performance criterion. However, this is not adequate in many situations if subsequent data on retention are to be obtained. So, the best we can do is to insist that we be alert to such situations and work out the appropriate solution for each experiment. For many of our studies on retention it was important when we discovered that retention was not related to ability level (16) so that possible confounding by subject differences becomes very unlikely. The same thing holds true regarding possible selective factors which may operate when long-term retention studies are used, and here I use long term in the sense that the subject leaves the laboratory after learning and returns at a later period for retention tests. Our conditions of learning may affect differential return of subjects. For example, Zeller (20) reports that subjects who were rather severely humiliated or insulted in his experiment had a much higher "no show" rate for subsequent retention tests than did those who were not so humiliated.

The above situations are quite obvious as illustrations of how the experimental situation may destroy the equality of groups. Let us turn to somewhat more subtle instances.

2. It is common in experiments using the white rat to form the groups by assigning subjects at random. Before assigning, restrictions may be placed on age, sex, weight, and litter so that these fac-

DESTRUCTION OF EQUIVALENT GROUPS AS A  
CONSEQUENCE OF RESEARCH PROCEDURES

Under this heading I want to consider a number of different ways by which the equivalence of our groups may be vitiated by experimental procedures which produce a loss of subjects. I shall consider enough samples so that we can be sensitized to the fact that these loss-of-subject situations may be varied and lethal. In general, the situations to be discussed have first used random or matched groups before the introduction of the experimental variables. We ask the question of what effect does the experimental treatment, producing loss of subjects, have on our results.

1. Suppose we were going to do a study on speed of rote learning as a function of meaningfulness of the materials to be learned. In the simple case we would construct lists which in so far as we can tell differ only on meaningfulness. Again, to keep it simple, let us consider only two levels of meaningfulness, hence two groups of subjects, one assigned to one list and one assigned to the other. Suppose we use a performance criterion, say, one perfect recitation. What we expect to find is that the list of low meaningfulness takes longer to learn than the list of high meaningfulness. Inevitably in such studies we find that some subjects will be unable to reach the criterion; they are unable to learn the list to which they were assigned. Let us assume that the first subject which comes to the laboratory is assigned to List 1, the second to List 2, the third to List 1, and so on. I think we would accept this as a method of assignment which should result in equivalent groups if factors such as time of day, experimenter, and so on, were equalized. When a subject fails to learn we assign the next subject to that list and proceed as if the subject had not been lost. We complete the experiment and discover that the two groups did not differ in terms of mean number of trials to learn the two lists. We might conclude that meaningfulness as manipulated here was not a significant variable. On second thought, even with this brief description, I am sure that no one would accept such a conclusion without additional data. In this particular case we would certainly express an interest in the number of subjects who were lost for failure to learn each list. If the number of subjects unable to learn each list was roughly the same we would

Let us assume that learning ability is related to socio-economic level. Let us further assume that the higher the socio-economic level the greater the longevity (due to better medical care). If these two assumptions have basis in fact (which they probably do) then the older the subject being measured the greater the selectivity toward better learning. The results obtained would probably not be the same that we would get if we took a sample of 5-year-olds and measured them for the next seventy years without loss of a single subject. Thus, the way the research has been done, natural selection may tend to favor more and more those who learn most rapidly. At the advanced age levels, we are dealing with a group that is superior in learning ability for that age as compared with a case where all people would have the same life expectancy. The implication is that if the samples obtained at the advanced ages were as representative of the entire socio-economic range as those at the younger ages, the decline in learning ability at the older ages would be much sharper than usually reported. I shall let you ponder the problem of how this research might be done to avoid the selection problem. But, if the reasoning above is correct, we do have a selection of subjects which throws a bias into the results.

4. I usually think it a good idea to analyze the data of a completed experiment in many different ways. One may make analyses on sub-groups within the general groups or fractionate the data in many other different ways to obtain all the information possible from a single experiment. This, in the long run of our science, is an economy move. For often an experiment will answer more questions or test more hypotheses than those for which it was specifically designed. Furthermore, the more angles from which one views a set of data the less likely it is that erroneous conclusions will eventuate. Often the final published report will contain only a small fraction of the analyses that have actually been made.

While I strongly recommend this multi-analysis approach, it cannot be so recommended without inserting some cautions. Sometimes in making these analyses we overlook the fact that in order for these internal comparisons to have substance they must conform to the sound procedures required of the over-all experiment. For example, if we are making a study of retention we might have data which would also allow us to see if there is a difference in retention of

tors are essentially equalized for all groups. In the particular study I am using for illustration (1), one of the experimental variables was the amount of weight on a Skinner-box lever. Thus, groups of rats had to work differentially hard in order to be rewarded. A common criterion of performance was imposed on all groups. It was discovered that the harder the rat had to work (more objectively, the bigger the weight he had to push) the less likely was he to reach the performance criterion. Hence, the loss of rats was directly related to size of weights. More and more rats were added until all groups had the same number of rats which had completed the task. But, it would appear that the groups now are no longer equal on all relevant variables. The heavier the weight the greater the likelihood that the rats remaining had greater strength or skill in bar pressing. If the investigator wished only to conclude that differences in original performance would result from differences in weight his evidence is overwhelmingly positive. However, in this case, extinction measures were taken and it is quite possible that these extinction measures represented a confounding of differences in ability and differences in the experimental variable.

3. A number of studies have been performed on learning as a function of age. There is opportunity in these studies for selection of subjects as a function of age so that the results may be biased, especially at the upper-age levels. A study investigating this relationship may be done about as follows. The investigator goes into a community and takes a sample of the people in various age ranges, say, 6-10, 11-15, 16-20, and so on. To each group he administers one or more learning tasks and then plots learning as a function of age. Results of such studies have been fairly consistent, namely, learning performance increases up to about the age of 20, remains fairly level to about 40, and then shows a very gradual decline. Of course, the shape of the curve varies as a function of the particular task but the relation indicated has some generality. Now it seems clear that the results are straightforward *as far as* the conclusion between age of those subjects used and learning is concerned. If the sampling from each age range is random, the curve is representative of the populations of each age range. However, I think there is a real doubt as to whether it represents only a relationship between age *per se* and learning. Let us see why this might be.

5. As I mentioned earlier experimental conditions which destroy equivalence of groups may take place in a great variety of ways. I cannot hope to cover the many ways in which this may occur in a specific experiment; so, with the discussion of one more type of research situation we will move along to other problems. In one study (5) a group of 50 boys and a group of 50 girls were asked to write down acts of destruction in which they had engaged. The results show that the percentage of boys writing down such acts was greater than the girls and that more acts were listed by the boys. In studying such results we must be careful that we do not conclude that boys are more destructive than girls. The only conclusion which is justified is that boys admit to more acts of destruction; whether boys are more or less destructive than girls cannot be concluded from such data.

Consider another situation faced by investigators in several experiments (e.g., 9). Words are flashed on a screen initially at an exposure time that is too brief to allow recognition. Gradually the exposure time is increased and the subject is instructed to report what word he perceives. When the exposure time is increased enough a point will be reached in which the subject can at least make a guess (based on partial cues) as to what the word is. According to the hypothesis which was involved in some of these experiments words with sexual meaning (e.g., "whore") will have a higher threshold (the exposure time would be longer) than neutral-toned words. The results of some studies have supported the expectation. However, we must be careful that the social aspects of the experiment do not produce these differences. It is quite possible that the sex words are not "repressed." Rather, it may be that the subject doesn't report such words until he is absolutely sure that he is observing correctly. This caution on the part of the subject is understandable if he wants to keep his embarrassment to a minimum.

#### BALANCING OF PROGRESSIVE ERRORS

I have been giving ways in which failure to get or maintain equivalent groups when manipulating a task or environmental variable may lead to confounding. If we are manipulating an environmental or task variable, we have said the groups must be equivalent on all



male *versus* female subjects. So, we divide our total group into subgroups of males and females with the intention of comparing retention. We should realize that in doing this these groups should not differ significantly on relevant variables for retention, e.g., degree of learning. Let us take another illustration which shows another facet of this problem.

Assume we did a study on spatial generalization (e.g., 3). The subject is faced with a row of seven lights. He is told that each time the center light comes on he is to press a key as quickly as he can. But at various times we light lights other than the center one to see if the subject responds. After a large number of trials we can plot the number of responses made to each light and this would be expected to show a decreasing frequency of response from the center light out on both sides, i.e., a gradient of spatial generalization. Another measure of generalization that might be used is the latency of response. More specifically it is expected that the latency gradient will be roughly a reciprocal of the frequency gradient, i.e., the more generalized the response the longer the latency or to say it another way, the fewer the responses the longer their latencies.

In such an experiment as this we might calculate the mean latency for responses to each light independently (using the number of responses as  $N$ ). If we did we would probably find that our expectation was not supported; that is, the more generalized responses might not have longer latencies than the less generalized responses. Indeed, the more generalized responses might even have shorter latencies. But, we would note that this was an inappropriate means of handling the latency data because we may have a subject selection process involved. Each subject is not represented at each light so that those subjects who did respond to the extreme lights may have very fast latencies, those who didn't may have very long latencies. So, when we calculate means based on all responses at each point (for each light) we are using subjects for the different points who have different "natural" latencies. The proper way to handle such data would be to use the latency for the center light (correct light) as a reference point and figure deviations from that for each subject for the generalized responses. In any event, we simply cannot get mean latencies for each position because of the subject selection which this method introduces.

each condition. Certain general statements can be made regarding factors which must be considered in making the choice.

1. I think there is only one situation in which the investigator *must* use the same subjects in all conditions. That situation is one in which he wants to investigate the influence of one or more conditions on behavior under subsequent conditions. Thus, if the investigator wants to study practice effects as such he must use the same subject at all stages of practice.

2. There are also a number of situations in which it would be ill-advised to use the same subjects in all conditions. These situations all fall under the general principle that if the effects of going from one condition, say A, to another condition, B, are different than in going from B to A (differential transfer) the same subject should not be used in both conditions (unless, of course, one wishes to study these differential effects as noted above). I have elsewhere (15) outlined specific situations in which such differential transfer is likely to occur so I will not repeat them here.

3. The same subjects are rarely if ever used in all conditions when a subject variable is being manipulated. To take an extreme case, it would be impossible to have the same subject be 8 years of age under one condition and 6 years of age under another, in that order. It would be nearly equally difficult to have a subject be an introvert under one condition and an extrovert under another.

4. There are a number of situations in which a study may be satisfactorily performed by either method; that is, by using a different group of subjects for each condition or by using a single group of subjects in all conditions. Under such circumstances the choice may be settled by reference to practical matters, such as number of subjects available, ability to get equivalent groups, amount of time available for each subject, availability of materials for several conditions, statistical analysis preferred, and so on.

Assuming that we have made the choice of using the same subjects in all conditions, I repeat that it is obligatory to balance out progressive errors. Several specific techniques are available for this. The essential principle on which these operate is that all conditions of the experiment must occur equally often at each stage of practice, all subjects considered. Again, since the details of these methods are

relevant subject variables. Now, if the same subjects are used in more than one condition when a task or environmental variable is being manipulated the same rule holds, namely, the subjects must not differ on relevant subject variables when different conditions are presented them. But it is a fact that as a result of having one or more experimental conditions the subjects *do* differ when presented a subsequent condition. There is no satisfactory way to prevent these subject changes; therefore, if the subject is to serve in more than one condition the experiment must be designed so that these changes in the subjects will not differentially influence conditions when the orders of conditions for all subjects are considered collectively. I call these changes in the subjects *progressive errors* and the method of handling them is some form of *counterbalancing*. Perhaps "progressive errors" is a misnomer; actually the term refers to the influence of behavior changes which occur as a consequence of continued experience with successive samples of the same class of materials or tasks. These behavior changes are usually said to be the result of *practice* and *fatigue* and are often referred to as *practice effects* and *fatigue effects*. It is an empirical fact that if a subject performs or practices on tasks which are relatively new to him, his performance will improve with continued practice. It is also a fact that sustained performance on a task may lead to decrements in performance and such decrements might be attributed to fatigue. In experimental work we can usually avoid any appreciable change attributable to fatigue by limiting the experimental time at any one session. It will be a rare situation, however, in which the investigator can say with confidence that there were no behavior changes attributable to practice. It is therefore extremely important that we recognize these experimentally irritating behavior changes which occur with successive practices; they will bias or distort the behavior which we wish to attribute to the manipulated variable unless our scheme of conditions is so arranged that the changes (whether increments or decrements) will fall equally on all conditions of the experiment.

In many kinds of research problems the investigator has a choice as to whether he will use the same subjects in all conditions of an experiment or whether he will use a different group of subjects for

extraneous stimulation on perceptual judgments. The judgment required was of verticality of a rod. This rod was luminescent and was presented to the subject in a dark room. The rod was sometimes tilted left, sometimes right, and the subject directed the experimenter to adjust the rod until he (the subject) judged it to be vertical. These adjustments were made under five different conditions. In one condition the subject was given a mild shock to the left neck muscle while directing the adjustments of the rod; in another he was given a shock to the right neck muscle. In a third condition he received mild auditory stimulation through the left ear while making the judgment and in a fourth condition the stimulation was in the right ear. Finally, the fifth condition was a control in which no stimulation was received while making the adjustments. (If these conditions seem a little peculiar let it be said that the particular theoretical orientation under which the investigation was done made them quite reasonable.)

There are five conditions. Each subject was given all five conditions, four trials under each. A trial consisted merely of one adjustment from a 30-degree tilt to a point where the subject reported the rod to be vertical. Half of the time the tilt was left and half time right so that any constant error in the judgments could not be attributed to a bias in the direction of movement of the rod. However, the report of this experiment contains the following with regard to the order of the five conditions:

The sequence of test conditions was: control, auditory-right, electrical-left, auditory-left, and electrical-right. (17, p. 342).

Note that with this order of conditions if there are progressive changes in behavior with successive trials they will influence the experimental conditions more than the control and the electrical-stimulation conditions more than the auditory-stimulation conditions. Among the several analyses made of the data obtained in this experiment was the comparison of the constant error under the control condition with the constant errors under the experimental condition, and the constant errors under each experimental condition with those under all others. Now we do not know whether progressive errors would be operative in this situation but it is a rare situation in which some sort of progressive change does not take place

available in other sources, I shall no more than mention them and make some brief evaluative statements.

1. Complete counterbalancing, in which each condition occurs equally often at each stage of practice and each condition precedes and follows all other conditions.

2. Some form of Latin square or systematic randomization, in which each condition occurs equally often at each stage of practice but all conditions do not precede and follow all other conditions.

3. Randomization, in which the order of conditions for each subject is assigned at random.

So far as I can tell, there is little to choose between the first two methods, except on practical grounds of number of subjects available and number the investigator wants to include. The number of subjects required for complete counterbalancing is  $r$  factorial, where  $r$  is the number of conditions. When we reach five conditions 120 subjects are needed, and with six, 720. With Latin squares or derivatives thereof the number of subjects is simply some multiple of the number of conditions.

The third method (randomization) probably should not be used when the number of subjects is small. One need not mortgage his soul to randomization in this case because the same effect as randomization can be achieved by systematically ordering the conditions by either of the first two methods. Strange as it may seem, randomization of conditions may be most easily justified when (as in many sensory experiments) data from a single subject constitute the entire data from a series of conditions. However, in these cases, the subject is given many trials under a single condition so that effectively it is as if many subjects were used and each given one trial on each condition. For example, if there are two conditions and 100 trials for each condition, randomization of the order of the 200 trials should result in the effective balancing of progressive errors. It would be equivalent (for balancing purposes) to giving 100 subjects a single trial on each condition in which the order of the two conditions was determined on a random basis.

So much by way of background. Turning now to specific research, I want to give you two published illustrations of the failure to balance progressive errors.

An experiment (17) was concerned with the effects of certain

The results obtained from these conditions were quite clear; as length of passage increased the per cent of words recalled correctly decreased, and as meaningfulness increased per cent recall increased. But it is quite evident from the design that the results are unmercifully confounded with practice effects, and, since all 32 lists were given in a single session, perhaps by fatigue effects. The results reflecting the length variable are probably more severely biased than those for meaningfulness. In the case of length, practice effects and increase-length effects would both increase throughout the course of the experiment. The 50-word lists would benefit much more from practice effects than would the 10-word lists. Very likely, therefore, differences in recall as a function of length are greater than observed here unless fatigue begins to influence the results late in the series. The variable of meaningfulness is not adequately balanced against progressive errors although it is in better shape than length of list. On the average the low-meaningful material (List 1) would appear earlier than List 2, 2 than 3, and so on. Therefore, if the progressive error is largely confined to practice effects, learning as a function of meaningfulness may be exaggerated over what would have been the case had the practice effects been equally balanced among the different levels of meaningfulness.

I think it can be seen that with 32 conditions (lists) and only 20 subjects perfect balancing cannot be obtained. But, with such a long series of conditions an imperfect balancing might be worked out which would be satisfactory without adding more subjects.

I think these illustrations make it clear that confounding may occur when progressive errors are not equally distributed over all conditions. This will be true whether the failure to balance occurs among the conditions for a given subject when the results for the different conditions for that subject are to be compared among themselves or when the failure to balance is among conditions for a group of subjects collectively. It will be true also where a task variable is being manipulated or where the influence of an environmental variable is being studied.

#### OTHER BALANCING SITUATIONS

In the discussion above I was concerned with a particular condition being behaviorally favored over another because of changes

with continued trials. If it did here we have the effects falling differentially on the conditions and we, therefore, do not know whether the differences among conditions as measured represent the effects of the conditions manipulated or the effects of these progressive changes or both. I think that in view of the direction of differences actually found the investigators might well argue that all differences could not be accounted for by progressive errors but it seems to me they would have a difficult time defending the proposition that the magnitude of the differences as found were not differentially influenced by progressive errors. However, no such defense would be necessary if the precaution of balancing the order of the conditions had been taken in the first place. In this case the balancing could have been accomplished within the conditions for each subject or among the 40 subjects which were used. This same failure to balance for progressive errors occurs in a second experiment by the same authors (19) but in a third experiment (18) the balancing is nicely accomplished by a partial balancing within a subject's conditions and a further balancing among subjects.

The second investigation (10) which I wish to discuss as an illustration of failure to balance progressive errors may be more serious than the above largely because the tasks used are known to produce large progressive changes in performance as a result of practice. The investigation was concerned with the retention of verbal lists as a function of two variables, namely, the length of list and degree of meaningfulness of items in the lists. Length of lists was varied four ways, namely, 10, 20, 30, and 50 words in a list. There were eight levels of meaningfulness to which I will refer with the numbers 1 through 8. In the procedure used, a list was presented to the subject, one word at a time, and immediately after the last word the subject wrote down as many words as he could remember. There were 20 subjects. All subjects learned and recalled all 32 lists (four lengths with eight levels of meaningfulness for each length). The order of presenting the lists, exactly the same for all 20 subjects, was as follows:

- 10 words long, all 8 lists in the order 1 through 8
- 20 words long, all 8 lists in the order 1 through 8
- 30 words long, all 8 lists in the order 1 through 8
- 50 words long, all 8 lists in the order 1 through 8

There are probably a number of reasons why experiments involving manipulation of environmental factors to discover the relationship with relatively permanent subject characteristics have been few in number. Undoubtedly, one factor is that they do require long periods since the assumption is that these relatively stable characteristics of subjects will not yield to short-term efforts. However that may be, in recent years some starts have been made on these problems. For example, in sensory deprivation studies, an animal may be raised in total darkness from birth to, say, one year and then tested for various visual skills or capacities. These measurements would then be compared with animals not so deprived. Such research presents no new general problems. So, therefore, it is the second paradigm, where subjects are initially separated on the basis of different amounts of a characteristic and then tested in a new situation, that I am concerned with in this section. Contrary to my procedure in the previous and subsequent sections, the discussion will be largely expository with little reference to actual research literature.

When we manipulate a subject variable, I think it will be recognized that we are dealing essentially with the same issues which arise when we wish to impute cause-and-effect status to events when using response correlation as an instrument of research. My approach to the problem will be somewhat different than that usually taken since I am interested in conclusions which can be reached when viewed in terms of the design of the research. Attempts to work out these research problems have been increasing directly with the increasing interest in clinical psychology where the immediate critical factors are subject variables. I shall run through some hypothetical problems to show the issues in as stark a manner as possible and then we shall search for solutions. But, I may say by way of preview that the analytical problem is such a difficult one that we can really only solve it by a series of interlocking researches dictated by working hypotheses. These working hypotheses in turn stem from the attempt to relate subject variables to fundamental behavioral phenomena. I use the word fundamental here in the sense of well-established phenomena and laws about them.

As a starting point let us consider a straightforward empirical study which relates the subject variable of chronological age to rigidity. (In a literary sense, high rigidity represents difficulty of



which take place in the subject's abilities as a consequence of the experimental treatments. There is a host of experimental situations in which we must deal with biases which the subjects bring to the situation. If these biases can favor one condition over the other and if we are not interested in the biasing effect, steps must be taken to eliminate these differential effects. Or, we might express this by saying that subject variables (biases) are not equal for two or more conditions of treatment and we want them to be equal. However, I find that these biases are usually involved in a confounding manner because of certain task characteristics. Therefore, I prefer to wait to discuss experimental errors resulting from these biases when I consider confoundings by task variables.

### CONFOUNDING BY SUBJECT VARIABLES WHEN MANIPULATING SUBJECT VARIABLES

In Chapter 2, I indicated two general ways by which subject variables may be investigated. First, we may manipulate conditions to discover what influence such manipulations have on specified subject characteristics. Thus, we might vary the amount of preschool training to see what influence this has on mental age. Secondly, we may choose groups of subjects who differ on some specified dimension (unitary or complex) and test for other differences in behavior.

The first method, that of determining causal factors lying behind individual differences on particular characteristics, actually involves the manipulation of environmental variables. Therefore, issues which have been discussed and others which will be discussed later concerning the manipulation of environmental variables will apply to this paradigm. If one wants to be really fussy one can say that the manipulation of an environmental variable is undertaken to discover the influence of the variable on subject capacities or skills. However, I have kept this research situation separate, not because it poses peculiar problems, but because it is concerned with modifications of relatively permanent skills and capacities and, furthermore, the research usually extends over long intervals of time. In contrast, the association one has when thinking of the typical manipulation of an environmental variable is a short-term study relatively unconcerned with modification of permanent skills of the subject.

and chronological age are nearly perfectly correlated. So now, we hold mental age constant and vary chronological age and we find no relationship. Now, our relationship is changed; chronological age is not the critical variable; mental age is. But what is there about differences in mental age which leads to differences in rigidity? We might suggest that the greater the mental age the faster the learning, and differences in rigidity are to be attributed to the fact that those people with higher mental age had stronger habits which were more difficult to overcome. Once again we go to the laboratory and proceed to equalize for strength of habits with varying mental ages. We find that the relationship still holds—the higher the mental age the greater the rigidity. We try again and suggest that differences are due to the rapidity with which learned habits are extinguished, with those with higher mental age showing slower extinction. We set up a new experiment to test this and indeed (to continue the fiction) we find that the higher the mental age the slower the extinction. In a sense, at this point, we have returned to our starting point. Essentially, we note that our operations used to measure rigidity and to measure extinction are much the same. We might have reached this point more quickly or more slowly than in my illustrations. But, what we are saying is that differences in rigidity represent differences in speed of extinction of responses and we may now proceed, to give further substance to this identification, by seeing if other variables known to influence rate of extinction now also influence rigidity in a like manner. If we get positive results in a series of such tests, we will arrive at a point where we will accept rigidity as being based on the more fundamental process of extinction. We may want to go further, depending on our theoretical predilections, but for our purposes the illustration is complete. I wish now to parallel this illustration with a different subject variable which makes the problem somewhat more difficult to handle. I shall also now emphasize the problems of experimental design since that is the essential reason we are getting involved in the matter of subject variables.

For this next problem let us say we want to work out the relationship between severity of schizophrenia and amount of rigidity. Assume that we can reliably sort schizophrenics into two groups which differ in severity. What are the problems we face in attempting to establish an unambiguous relationship between severity of

changing habits, low rigidity quite the opposite; in the actual experimental situation several different techniques have been used to give rigidity operational meaning.) To do the research we sample different chronological ages and get measures of rigidity for the sample at each age. Suppose we found a positive linear relationship between age and rigidity. What do we conclude besides the fact that age and rigidity are directly and linearly related? Perhaps we would not care to conclude anything else; after all, we have laid bare a relationship and science is a search for relationships. Of course, if we do obtain the relationship as indicated we have a new way to measure or diagnose age. That is, knowing the rigidity score for a person we can tell his chronological age. The error in our estimation of chronological age based on rigidity scores may be somewhat greater than we would obtain by examining birth records or asking people what their ages are, but nevertheless we can predict age from rigidity scores. If I seem facetious it is only to emphasize the weak nature of the conclusion at which we have arrived by our single piece of research. As scientists engaged in a perpetual attempt to reduce relationships to basic cause-effect relationships, we would be quite discontented with stopping our research at this point where all we know is that there is a positive relationship between rigidity and chronological age. We know that chronological age is merely a convenient dimension of time and we must look for other changes which occur with time. But let us get away from this illustration for a moment to show that it is not an artificial one.

One of the few substantial findings in the area of problem-solving is that men do better on such problems than women. No scientist that I know of is content with this finding; rather, it raises the problem as to why men are better. Some might suggest that there is some genetically based difference which leads to the behavioral difference and then set about to search for this genetic differential. Others may attempt to relate this to experiential differences, thus relating it to a process about which we already know. The important point is that differences in behavior related to subject variables only start research, for in the typical case these differences must be related to more fundamental behavioral processes.

Let us turn back to the rigidity illustration. Having related rigidity and chronological age, we note that up to a certain point mental age

white blood count, weight, learning ability, number of siblings, and on and on. The point is that since we do not and cannot possibly know all the variables influencing rigidity and since we want all these variables equalized for the groups as long as they are not criteria used in diagnosis, we reach an impasse. How can we extricate ourselves from this onerous research situation so as to salvage something with scientific respectability?

We noted above the difficulty of the matching problem. Let us make an outlandish assumption to see what its implication is. Suppose we were able to match our two groups on every variable (still excepting, of course, degree of schizophrenia). Now we test for differences in rigidity and we find that the two groups do indeed differ in rigidity as expected. We might feel particularly smug with ourselves, until we seek the implication of the finding. For when we do we discover that all we have done is add another diagnostic criterion to our diagnostic armamentarium; we have added that and no more. If the rigidity tests separate clearly it may be of considerable benefit for future diagnosis—and I do not minimize the practical importance of the finding—but it is at best only a first step in an analytical framework which must be built up to give scientific meaning to the relationship. And of course, we have obtained this result under conditions which could not possibly exist. I have brought this matter up in this way to try and show that what seemed to be a desirable goal—matching on everything—actually allowed us only a pitifully weak conclusion when viewed as an analytical step in science. We must turn to another alternative.

As I view this situation escape comes only by the use of some form of theoretical approach. The matter of theory is to be discussed in later chapters but there is no way to avoid a little of it now. Turn back to the research situation of the schizophrenic-rigidity problem as I first outlined it. I said that the diagnostic criteria must not include rigidity. The problem arose because certain implications of the schizophrenic syndrome led to the proposition that rigidity would be different for different levels of severity of illness. This is theory-like language. It says that if this is so, then this must be so. Even if I were able, I would not detail what premise (or premises) might be stated which led to the deduction that rigidity would vary as a function of severity. But, in investigations in which subject variables

schizophrenia and rigidity? (I shall not consider problems peculiar to this illustration such as the matter of making contact with the severe cases, since I am using this problem only as an illustration of a general design issue.) First we must satisfy ourselves that a difference in rigidity was not one of the criteria used in making the diagnoses of severity of illness. For, if this is true, and we find differences in rigidity on our experimental task, all we can say is that performance on the experimental task confirms the reliability of the diagnosis and perhaps adds a little to our idea of the generality of rigidity in the individual. (This situation parallels the illustration given in the previous chapter concerning "effect" of length of time spent in Boy Scouts on community adjustment.) Without going into any detail, what we want to have is a working hypothesis concerning the relationship between degree of schizophrenia and rigidity suggested by the implications of the syndrome but where the degree of rigidity was not used to sort out the two groups. This latter is quite possible if the diagnostician can clearly specify the criteria used in making the sorts on severity.

The second design problem is more difficult to solve satisfactorily. We want the two groups to differ only on severity of schizophrenia. How do we accomplish this? When manipulating task or environmental variables we have (as one technique) used random assignment to accomplish the equalization of factors other than the one we are varying. We might think at first glance that we could do the same here. We could take a random sample of each of two large groups (one diagnosed as mild the other severe) on the assumption that this would equalize for other factors. Very quickly, however, a second glance will show us that this could be a lethal procedure. Suppose the two groups differed in age with the severe group having an older mean age. Suppose further that we actually knew that rigidity and age were positively related. If we find a difference in rigidity between our two schizophrenic groups we would quickly realize that this could well be independent of the severity of schizophrenia. So, what do we do?

It would seem that we would have to turn to a matching procedure. So, we first match on chronological age. But this only? Well, no, perhaps we should match on sex, mental age, socio-economic background, racial background, education, length of stay in hospital,

with their idea of *construct validity*, although the reference to fundamental processes (independently derived) is not made explicit. Furthermore, I receive the distinct impression that if we had techniques for factor analysis of performance scores which are not linearly related, this factor analysis would remove the need for an independent idea of construct validity. But, I cannot be sure, for they speak of theory which must be traced to behavioral measures. The vagueness comes in their failure to specify the nature of these behavioral measures and this, according to the argument I have advanced, makes considerable difference.

If I stopped at this point I would be in the position of essentially asserting that a theoretical approach has solved a design problem. To a certain extent this is true, but it is by no means entirely true. You will remember that the design problem arose because we could not match on all possible relevant variables. Our theoretical approach doesn't handle this directly; that is why we must insist that before we begin to take any working hypothesis seriously we must have a series of different tests of its implications. This is one area of investigation where a single piece of research by itself is of little or no scientific value. Let us see why this is so and how we work out such a problem. Let us use the subject variable of anxiety since that is a familiar one in the literature at the moment. We separate two or more groups on the basis of a pencil-and-paper test, these groups differing in anxiety. Anxiety is conceived as a drive and the experiments are designed to see if differences in anxiety (conceptualized as differences in drive) lead to predictable results, the predictions being made on the basis of what is already known about how drives operate to influence behavior plus any theory which has been built up around these facts. (I am not particularly concerned about the success or failure of the identification of drive and anxiety; I am concerned with it only as an illustration of the nature of an approach to be taken in fitting subject variables into our empirical structure of cause-effect relationships.)

Suppose that as a first test we say that if anxiety is a drive, differences in performance in a conditioning situation should be found between two levels of anxiety. We enter the laboratory and find indeed that the performance differs in the expected direction. If we stop at this point we are subject to all sorts of criticisms. In the first

have been attacked with some degree of success, certain working hypotheses have been first advanced and then the implications of these hypotheses explored by research. More specifically, these working hypotheses *relate the subject variables to fundamental behavioral processes*. Again let me say by fundamental behavioral processes I mean processes which have been investigated independently, for which there are laws with environmental variables, and for which there is evidence that they permeate a wide range of behavior phenomena. I would include under this idea of fundamental such processes as maturation, learning, forgetting, inhibitory processes, motivation, and so on.

The investigator starts out with an idea or hypothesis that differences in subject variables (such as differences in severity of schizophrenia) reflect the operation of more fundamental processes. What he says, in effect, is, if this difference is due to this or that process (or a combination) then he would expect this (difference in rigidity) to obtain. He is applying a set of principles of behavior about which we already know considerable to another area of behavior other than that used to derive the principles in the first place. If a reasonable number of tests of the implications of the application are positive we then begin to accept the original hypothesis which identified differences in a subject variable with a difference in a more fundamental process.

The studies on manifest anxiety, originally stemming from Iowa, have taken this approach. The fundamental hypothesis advanced was that differences in anxiety represent differences in drive. If this is so, then according to what was known about drives and their theoretical elaboration, such-and-such should happen in certain situations. This is also the approach taken by Eysenck (e.g., 8). Eysenck first (by factor analysis) obtains what to him are general descriptive dimensions of personality. Then he asks himself what fundamental processes lie behind or cause individual differences on these dimensions. For one dimension he suggests that differences in inhibition, as inferred from classical experimental research, may be involved. He then proceeds to test for differences in magnitude of empirical phenomena (believed to reflect differences in inhibition) for individuals who score differently on his descriptive personality dimension. Cronbach and Meehl (6), if I understand them, almost reach this position

if we search for more basic explanations, when drives in general are explained. And it seems to me that at the same time we have largely solved our design problem, a problem which could not be solved without recourse to the relating of this subject variable to a fundamental process of behavior.

Now it is quite possible, indeed likely, that certain subject variables cannot be reduced to manifestations of something about which we already know. After all, our science is relatively young and there may well be what I have called fundamental processes which are as yet not discovered or perhaps poorly understood. I see no easy escape from this situation. If we cannot relate differences in subject variables to differences in known processes which lie behind developmental stages—which lie behind individual differences of all kinds—then we may be in a position to postulate a new process, but the experimental validation of this process will be slow in finding acceptance because of difficulties of the research design which I have discussed. The postulation of a new process does not abrogate these design responsibilities.

There are two additional points regarding this research process on subject variables which should be mentioned at this point. Again I will use anxiety to preserve continuity. Suppose that in the initial stages of research some one objects to the interpretation of the results on the grounds that perhaps anxiety and general learning ability are related and that differences in performance tentatively attributed to anxiety may actually be due to differences in learning ability. In some cases, depending on the nature of the theory, such possibilities can be effectively eliminated by differential predictions for different learning situations. Thus, in the case of anxiety as identified with drive, predictions may be made that for one kind of learning task high-anxious subjects will do better than low, while for a different task the prediction may be reversed. Confirmation of such predictions would pretty much destroy the idea of a correlation between general learning ability and anxiety.

The second point is that in the case of certain subject variables, attaining the objectives of the line of research outlined for anxiety may be greatly facilitated by another approach. This facilitation is produced by introducing experimental conditions which change the amount of the characteristic for subjects who originally did not



place we have arrived at what I have called previously the pitifully weak conclusion that we now have a new technique—performance in a conditioning situation—to diagnose anxiety. Secondly, and this is far more important at this stage, we are liable to the criticism that anxiety may not be the critical variable involved. It is like the relationship between chronological age and rigidity. Perhaps the groups differing in anxiety also differ in age, intelligence, learning ability, or in a great many variables that might influence performance in the conditioning situation independent of anxiety. Basically we are at the same vulnerable and helpless point that we were in our study of schizophrenia and rigidity. How do we get our groups equivalent on all variables except anxiety? The answer is we don't. We might match on factors which are known to influence performance in the conditioning situation and perhaps on other factors if it is convenient to do so. Of course, if we do accomplish such matching and suddenly the relationship between performance and anxiety disappears we have identified the critical variable as something other than anxiety. The whole pattern of research would change at this point. But, let us continue the illustration by assuming that matching leads to no change in our results—performance is still related to anxiety. (It may be noted parenthetically that by such matching procedures we are obtaining a great deal of information about what subject variables are *irrelevant* to performance in a conditioning situation.) But we have asserted and still must assert that we cannot be confident that we have eliminated all variables as possibly more basic to the relationship than anxiety. That is, some variable which is somewhat correlated with our measure of anxiety may still be responsible for our empirical relationship so that if we knew what this variable was and held it constant while still varying anxiety our relationship would disappear. There is no completely satisfactory solution to this dilemma. But what we do is push the implications of the drive hypothesis through a series of experiments, exploring many possible implications. If the results rather consistently parallel those obtained when other drives are manipulated our confidence is substantially increased that we are justified in relating the results obtained from this variable (anxiety) to our body of knowledge concerning other drives. We thus gradually remove anxiety as a behavioral phenomenon requiring independent theoretical solution. It will be explained,

of any variable. The possibility looms large, then, that specious identification may occur. To decrease such a possibility, I have suggested that many confirming tests must be available. Furthermore, the more comparable the precise laws resulting from the manipulation of the subject variable (e.g., anxiety) and other drives (e.g., thirst) the greater the confidence in the unification. A reasonable amount of comparability among the laws will at least allow one to include the subject variable in the class of operations under which all drives are placed. That the exact laws will not be the same for all drives will certainly be expected and when they are different the operational distinctions within the class are to be maintained. I shall develop this matter more fully in a later chapter.

It seems to me, all matters considered, that until some more fruitful way of dealing with subject variables is advanced, the above procedure should be tried in spite of its dangers. The procedure has the potentiality of solving a very difficult methodological problem and is at least worth a serious try for several subject variables. I do not see how such research can make us "worse off" than we now are in our dealings with subject variables and it might lead to real progress.

This discussion on manipulation of subject variables has been long and involved. Let me summarize the points which I think are useful in guiding our thinking about this difficult research paradigm.

1. If we separate subjects on a particular characteristic or cluster of characteristics and then if we test them on one of the same characteristics, we are only demonstrating the reliability of our original selection (providing our test gives commensurate differences). It will add a little to the generality of the trait involved if the diagnostic tool and the subsequent test are different in specifiable ways. The fact that in this situation we only demonstrate the reliability of our diagnosis may seem simple-minded when stated so starkly, but it is not so obvious when the vagaries of clinical diagnosis are involved.

2. If we separate subjects on a particular characteristic or cluster of characteristics, and then test them on an aspect of behavior *not* used in the original separation, the following points are relevant:

- (a) From the design standpoint, the groups should be equated on all other variables other than those used to effect the separation. Random selection cannot accomplish this. Matching on variables

differ on the characteristic. For example, in the case of anxiety, we might choose two random groups from the same population and attempt to experimentally induce high anxiety in one group and low in another to see if our results on some task confirm results based on selecting different groups having more or less permanent differences in anxiety. If the results do conform, we have in a sense eliminated the possibility that our results obtained from permanent anxiety groups could have been a function of some unknown factor being partially correlated with anxiety. Perhaps two or three tests would be necessary to be confident of this conclusion but it would shorten our program of research considerably if the results were positive. If the results are negative only slight doubt is cast on the original hypothesis. For, our experimental conditions designed to introduce differences in anxiety may be inadequate or, if judged adequate, the experimental anxiety may not serve as a drive in the same sense as the anxiety "naturally" present in different amounts in subjects. Nevertheless, we should always examine the situation to see if it is possible that we might experimentally introduce difference in subject variables. Obviously, there are many subject variables in which this is not possible. We would probably find it difficult to experimentally change the chronological age of our subjects and there would probably be objections from certain quarters if we tried to experimentally induce different degrees of schizophrenia in originally normal subjects. Such situations sometimes lead to work on lower animals, which may then be used to support inferences on the human level.

In the above discussion I have indicated that before the identification of a subject variable and a basic behavioral process can be made, several confirming tests must occur. I have also suggested that confidence in the identification is increased if relatively unique predictions can be made and confirmed. Thus, in the case of anxiety, the expectation (based on drive theory) is that in a situation where there is little interference high-anxiety subjects will be superior in performance to low-anxiety subjects but the reverse will be true if the interference is high. Nevertheless, this whole identification process has a danger component in it. Behavior can only change in an upward or downward direction as a consequence of manipulation

with any assurance that this relationship is between our manipulated variable *per se* and our other measure because of the very real possibility that some variable co-varies with our specified variable and that this co-variate is "really" responsible for our obtained relationship. But, supposing we get negative results? Supposing that different degrees of schizophrenia bears no relationship to rigidity. With such a finding I think we would find that a preponderant majority of psychologists would conclude that the hypothesis must have been incorrect, i.e., the conclusion is that there is no relationship between degree of schizophrenia *per se* and rigidity. On strictly logical grounds I do not see how this conclusion can be arrived at with any more confidence than in the case of a positive conclusion as discussed earlier. Isn't it reasonable to suggest that a subject variable which is related to degree of schizophrenia is *negatively* related to rigidity and that the effect of this variable counteracted the effect of schizophrenia so that no relationship was measured? Thus, if this foreign variable were held constant the relationship between schizophrenia and rigidity would emerge.

In spite of this argument I suspect that there are grounds for the more ready acceptance of negative results than positive results in research of this kind. Correlation matrices involving many, many dimensions of subject capacities or skills rarely show strong negative correlations. Furthermore, I suspect one could argue that, all factors considered, we have to have a fairly delicate balance between a negative and positive effect in order to produce a zero relationship and that the probabilities of this occurring are low. I suppose also one might argue at a somewhat general level that contrary relationships among subject variables and a particular performance could be opposed to evolutionary survival theory. In any event, as I said earlier, if I perceive our scientific mores correctly, negative results in manipulating a subject variable are more quickly accepted than positive results, if the later results are initially stated to indicate a fundamental relationship. I think there is some justification for this differential acceptance. My only caution is that we may accept negative results a little more readily than the logic of the situation justifies.

will also not solve the problem, for potentially we should have to match on every possible variable known. The matching problem is not solvable but even if it were we would arrive at a weak conclusion scientifically, namely, that if we get a difference in performance we have discovered another diagnostic tool.

(b) Since we cannot match on everything, if we proceed with the research anyhow we are running a real risk that differences which we may obtain are a function of a partially correlated subject variable which we have not identified.

3. The best solution seems to be to not try to match except on variables which are known to be relevant to the task or others which if not known to be relevant, are easily handled in a matching situation. Then, we attempt to relate or identify the subject variable under investigation to a more fundamental behavior characteristic about which we already have considerable laws and see if our expected relationship holds for this subject variable. A single experiment is relatively worthless in this context. We normally would expect a number of positive tests before we feel confident that our original hypothesized identification is tenable.

4. For some subject variables this research may be accelerated by the use of experimental manipulations which temporarily induce different amounts of the subject characteristic.

In completing this discussion of the problems involved in manipulating subject variables, I want to consider the implications of negative results in this type of research. When manipulating an environmental or a task variable, whether or not the results *per se* are positive or negative with regard to the variable is a matter that has little bearing on a judgment whether or not the design of the investigation was sound. That is, if I manipulate an environmental variable and get a positive relationship between my variable and behavior, I accept this relationship only if I can perceive that no confounding variable has operated. If I get negative results i.e., no relationship between the variable and behavior, I would likewise accept the results if I can perceive no confounding.

When a subject variable is manipulated the major problem, as discussed extensively above, is what to make of positive results. I have said that when a positive result is obtained we do not know

18. WAPNER, S., WERNER, H., & MORANT, R. B. Experiments on sensory-tonic field theory of perception: III. Effect of body rotation on the visual perception of verticality. *J. exp. Psychol.*, 1951, 42, 351-357.
19. WERNER, H., WAPNER, S., & CHANDLER, K. A. Experiments on sensory-tonic field theory of perception: II. Effect of supported and unsupported tilt of the body on the visual perception of verticality. *J. exp. Psychol.*, 1951, 42, 346-350.
20. ZELLER, A. F. An experimental analogue of repression: III. The effect of induced failure and success on memory measured by recall. *J. exp. Psychol.*, 1951, 42, 32-38.

## REFERENCES

1. APPLEZWEIG, M. H. Response potential as a function of effort. *J. comp. physiol. Psychol.*, 1951, 44, 225-235.
2. BECK, S. J. The science of personality: Nomothetic or idiographic? *Psychol. Rev.*, 1953, 60, 353-359.
3. BROWN, J. S., BILODEAU, E. A., & BARON, M. R. Bidirectional gradients in the strength of a generalized voluntary response to stimuli on a visual-spatial dimension. *J. exp. Psychol.*, 1951, 41, 52-61.
4. CHAPIN, F. S. *Experimental designs in sociological research*. New York: Harper, 1947.
5. CLARK, W. H. Sex differences and motivations in the urge to destroy. *J. soc. Psychol.*, 1952, 36, 167-177.
6. CRONBACH, L., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, 52, 281-302.
7. DU MAS, F. M. Science and the single case. *Psychol. Rep.*, 1955, 1, 65-76.
8. EYSENCK, H. J. Cortical inhibition, figural aftereffect, and theory of personality. *J. abnorm. soc. Psychol.*, 1955, 51, 94-106.
9. MCGINNIES, E. M. Emotionality and perceptual defense. *Psychol. Rev.* 1949, 56, 244-251.
10. MILLER, G. A. & SELFRIDGE, J. A. Verbal context and the recall of meaningful material. *Amer. J. Psychol.*, 1950, 63, 176-186.
11. OSEAS, L., & UNOERWOOD, B. J. Studies of distributed practice: V. Learning and retention of concepts. *J. exp. Psychol.*, 1952, 43, 143-148.
12. POSTMAN, L., & ALPER, T. G. Retroactive inhibition as a function of the time of interpolation of the inhibitor between learning and recall. *Amer. J. Psychol.*, 1946, 59, 439-449.
13. ROSENZWEIG, S. Idiodynamics in personality theory with special reference to projective methods. *Psychol. Rev.*, 1951, 58, 213-223.
14. SEEMAN, W., & GALANTER, E. Objectivity in systematic and "idiodynamic" psychology. *Psychol. Rev.*, 1952, 59, 285-289.
15. UNDERWOOD, B. J. *Experimental psychology*. New York: Appleton-Century-Crofts, 1949.
16. UNDERWOOD, B. J. Speed of learning and amount retained: A consideration of methodology. *Psychol. Bull.*, 1954, 51, 276-282.
17. WAPNER, S. WERNER, H., & CHANDLER, K. A. Experiments on sensory-tonic field theory of perception: I. Effect of extraneous stimulation on the visual perception of verticality. *J. exp. Psychol.*, 1951, 42, 341-345.

reference to other conditions which have greater meaningfulness; the subject is still given positive treatment along the same dimension that those with higher degrees of meaningfulness are treated. If we varied the size of the target on a pursuit rotor, zero size would not make a very sensible task for the subject; nor would presenting zero frequency of sound waves to a subject make much sense unless we were checking on false responses when measuring pitch thresholds. Thus, we have reference conditions but these will rarely involve a "true zero amount" of the task variable being manipulated. The same situation obtains when we are manipulating a subject variable. We would rarely if ever have zero amount of a given characteristic for one group of subjects. Again, we might have a basic reference group which had a greater or lesser amount of a characteristic than did the other groups used but it is difficult to conceive of a situation in which we would have zero amount.

The implication of these distinctions is that when we are manipulating an environmental variable and where this variable is under discussion because it may be confounded by another environmental variable, the confounding usually results from failure to recognize the need for a control group or groups. The control-group issue is of little consequence in any other research situation, but is the critical issue in this situation. Therefore, the entire discussion in this section (and it is a long section) revolves around the use of control groups. All phenomena which result from the manipulating of an environmental variable are defined by the E/C, S-R type of definitions as presented in Chapter 3. The simplest sort of an experiment is represented by E/C operations. Or, to state this more positively, we usually don't have an experiment unless we have two conditions and when an environmental variable is involved these are commonly an experimental and a control group.

In the E/C operations the control does not, of course, have to be a separate group of subjects; the same subjects may serve in all conditions as discussed in the previous chapter when dealing with equivalence of groups on subject variables. In the present chapter, however, I shall, for simplicity, refer to control groups with the understanding that this includes control conditions if the same subject is used in all conditions. Because of the very close tie between design problems and definitional problems in the case of E/C oper-



## *Research Design: II*

### CONFOUNDING BY ENVIRONMENTAL VARIABLES WHEN MANIPULATING ENVIRONMENTAL VARIABLES

In the previous chapter, I discussed confoundings by subject variables when manipulating environmental, task, and subject variables. These confoundings are identified with cells 3, 6, and 9, respectively, in the table on page 91 of the previous chapter. I now want to consider cell 1 of this table. This cell refers to confoundings by environmental variables when manipulating environmental variables. The treatment of this topic involves a little preparation by way of indicating somewhat more specifically the nature of the problems which arise.

The problems centered around use of control groups arise almost exclusively in research where an environmental variable is being manipulated. The pure case of the control group is one in which the subjects of this group have not been given any experimental treatment; their behavior is then compared with that of another group (the experimental group) which *has* been given experimental treatment. To be unnecessarily precise I suppose we must say that you can't give a control group zero treatment; the subjects in this group do not exist in a vacuum while the experimental group is being treated. But in a practical sense the control group does receive a zero amount of the environmental variable when the effect of such a variable is being studied. The point I wish to make is that in contrast to this situation, when we are manipulating a task variable we rarely have a group which is given zero amount of treatment. Some examples may shape the contrast. If we are manipulating meaningfulness we don't have material which has zero meaningfulness. If a condition is called zero meaningfulness it is called this only with

## *Research Design: II*

### CONFOUNDING BY ENVIRONMENTAL VARIABLES WHEN MANIPULATING ENVIRONMENTAL VARIABLES

In the previous chapter, I discussed confoundings by subject variables when manipulating environmental, task, and subject variables. These confoundings are identified with cells 3, 6, and 9, respectively, in the table on page 91 of the previous chapter. I now want to consider cell 1 of this table. This cell refers to confoundings by environmental variables when manipulating environmental variables. The treatment of this topic involves a little preparation by way of indicating somewhat more specifically the nature of the problems which arise.

The problems centered around use of control groups arise almost exclusively in research where an environmental variable is being manipulated. The pure case of the control group is one in which the subjects of this group have not been given any experimental treatment; their behavior is then compared with that of another group (the experimental group) which *has* been given experimental treatment. To be unnecessarily precise I suppose we must say that you can't give a control group zero treatment; the subjects in this group do not exist in a vacuum while the experimental group is being treated. But in a practical sense the control group does receive a zero amount of the environmental variable when the effect of such a variable is being studied. The point I wish to make is that in contrast to this situation, when we are manipulating a task variable we rarely have a group which is given zero amount of treatment. Some examples may shape the contrast. If we are manipulating meaningfulness we don't have material which has zero meaningfulness. If a condition is called zero meaningfulness it is called this only with

report clearly implies that the change in hostility resulted from group therapy.

I think you will agree that this experiment is quite meaningless without at least one control group. The change in hostility scores might have taken place with no group therapy. There is a clear discrepancy between what was concluded and what can legitimately be concluded. In evaluating such experiments one need go no further than this in the analysis. However, one can often perceive other environmental factors which could have produced the same change as the one attributed to the independent variable. In the present experiment, for example, the subjects came to the University of Chicago for the special training session. If the general liberal tradition involving attitudes toward different ethnic groups is in fact exemplified at this university then the changes observed may have been due to the assimilation of this tradition and not to the therapy. While in this case it may be possible to think of factors which would produce the change (in addition to therapy), one is not obligated to do this in order to reject the conclusions drawn in this research. Changes in ethnic hostility may take place for reasons we cannot identify; the only way to handle this possibility is to give a control group the two testings without the 35 hours of therapy. Only then can we attribute differences to the influence of the therapy. From the experiment as it now stands it is impossible to conclude anything more substantial than that the scores on the second test showed less hostility than those on the first. The causal condition for the change cannot be specified. By way of looking forward, I might say that other issues concerning analyses of experiments of this type will be discussed later.

2. In discussing the operational definition of reminiscence in the previous chapter, I indicated that a number of investigators carried out research on this topic without inserting a control group. The basic procedure was to give the subject a number of learning trials, then an immediate retention test, followed by a second retention test after an hour, 24 hours, a week, or whatever interval with which the investigator was concerned. If the outcome of the second retention test was superior to the first, reminiscence was said to have occurred. The implication was that "something" happened between the first and second retention test that enhanced the recall. This "something"

ations, you will have to tolerate some repetition between the present exposition and the one given when discussing E/C definitions.

#### FAILURE TO USE A CONTROL GROUP

Boring (6) has reviewed the history of the use of control groups in psychology. He points out that the first use of a bona fide control group was the classical Thorndike-Woodworth study on transfer of training in 1901. Since that time there has been a gradual increase in the frequency with which control observations are used. This increase might be accounted for because of shifts in research areas in which the phenomena require control groups for adequate definition. But it also might be expected because of the progressively more analytical nature of a science as more and more phenomena are discovered, and as stimulus complexes are broken down into more unitary dimensions. Whatever the historical correlates are to this trend, at our present stage of methodological sophistication regarding the necessity for the use of a control group when establishing the reliability of a phenomenon based on environmental manipulation, it is discouraging to find reports in recent literature where there is complete failure to use a control group of any kind. Let us be sure we understand the seriousness of this situation as disclosed in actual research reports.

1. The purpose of one study, as given in the introduction to the report, was to determine the effects of a series of group therapy meetings on ethnic hostility (20). It is not our concern here whether or not the instrument used to measure hostility was adequate. The principal points of the procedure were simple. The 24 subjects were first measured on an ethnic hostility scale. Then, for six weeks they participated in a client-centered counseling training program. As a part of the program group therapy sessions were held with a trained therapist in charge. The total time of such sessions was about 35 hours and this was the manipulated variable of the research. At the end of the six-week period the subjects were again measured on hostility by the same instrument used at the initiation of the program. A comparison of the initial and final scores shows that subjects exhibited less ethnic hostility on the final test than on the first. The

4. For some reasons, investigators using various procedures designed to reduce the distress of the mentally ill have been particularly myopic toward the use of a basic control group to determine if the therapeutic procedures produce positive results. This has been true in some cases where electroshock has been used (e.g., 7) and where frontal lobe operations are employed (e.g., 12). We shall later see an illustration of the use of an inappropriate control group in lobotomy research. When we attempt to evaluate the influence of face-to-face talking therapy (e.g., 11), we should by now realize that no conclusions concerning the effect of therapy *per se* can be achieved unless appropriate control group or groups are used. I do not think our literature should be cluttered with these anachronistic procedures even if editors and authors recognize and publicly admit these shortcomings, for the shortcomings are fatal in a scientific sense.

I do not wish to spend more time on this matter; the failure to use a control group is such an obvious error that our exposition of it should not be extended. We have much more ground to cover and the errors yet to come are for the most part more subtle than the simple failure to use a control group. But, it may be asked if there is any time or any situation in which a control group is *not* needed to establish the relevance or nonrelevance of a single change in a condition. Undoubtedly there are such situations, although I have been unable to think of any in which only one or two observations of behavior are to be made. However, let us take a couple of situations in which *many* observations are recorded and see if you would not agree that an error in our conclusion is unlikely even though we have failed to include a control group.

Suppose we have a group of adults who by standard testing procedures have been classified as imbeciles for 20 years. Each year for 20 years they have been tested and each year the test record shows no appreciable change. Then, a new drug ("anti-imbecile") is placed on the market. On one day all members of the group are given an injection of the drug and the next day the test scores all fall within the normal range. Although it is remotely possible that something other than the drug caused the change, it is highly unlikely that anyone would care to defend this position strongly. The 20 years of

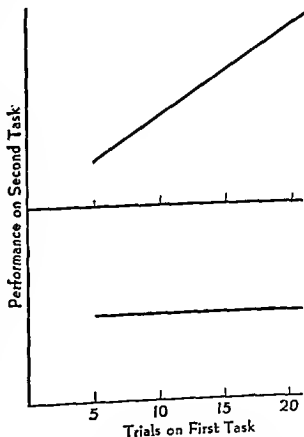
was not referred to as an environmental factor but rather as something spontaneously occurring within the subject. But, as was subsequently pointed out by other investigators, the first retention test may have served as a learning trial; thus, the degree of original learning was not the same for both retention tests. Time and degree of learning are being confounded. Obviously a control group which took the second test immediately after the first was needed in order to determine whether or not a pseudophenomenon was involved.

3. Next, let us consider a study in the area of frustration (4). The basic design of the experiment followed the order: test, period of blocking, test again. The subjects were children and a special measure of behavior called *constructiveness of play* was carefully worked out. Essentially the measure represents maturity of play and is quite highly related to intelligence. In the first observation period, the subjects were rated for constructiveness during a 30-minute free-play period. Then a period designed to induce frustration was inserted. Following this the subjects were put back into the original play situation and again constructiveness of play was measured. The results show that the average constructiveness age of play in the second test period was considerably below that of the first and it was concluded that this resulted from the treatment designed to induce frustration (a more complete analysis and criticism of this experiment has been published, 10).

To my way of thinking a control group is mandatory for such an experiment, this group having the two test periods but not having the blocking. It may well be that constructiveness of play in the second period would be considerably reduced even without the period of blocking. (It is interesting to note that the investigators in this research have, for other research situations, used an explanatory concept of *satiation*. This concept would roughly predict a decrease in constructiveness merely as a result of continual exposure to the same situation.) Note that we cannot say that blocking had no effect on constructiveness but neither can we say that it did. More precisely, in line with the previous chapter, we would say that frustration had not been demonstrated, hence was not defined. This is the intolerably ambiguous situation which use of a control group would obviate.

## SYSTEMATIC VARIATION AND THE CONTROL GROUP

In some research, conclusions of importance can be reached even if no control condition is used provided the experiments involve manipulation of a variable so that it is tested at several points. As an example, suppose that I performed an experiment on transfer of training in which degree or amount of first-task learning is varied by several steps, say, 5, 10, 15, and 20 presentations of the first task. All subjects are given the same number of trials on the second task and my basic measure of transfer is performance on the second task. Note that I do not have a control group, that is, a group which had *no* trials on the first task. Assume we have completed the experiment and obtain the results indicated in the upper part of the accompanying figure. I can conclude with confidence (merely reflecting the graph) that for these materials performance on the second task increases directly in proportion to number of trials on the first task. If I equate performance directly with transfer I can say that transfer increases directly as number of trials on the first task increases. But, strictly speaking, I do not know whether my measurements on the second task reflect increasing amounts of positive transfer or decreasing amounts of negative transfer or some combination. If I had a control group which had no trials on the first task I could determine the



observation essentially provide the control condition. One might suggest that to be on the safe side an injection of a harmless liquid should have been given to some of the members as a control but I for one would not push this. Such patients have had plenty of opportunities to receive placebos of all kinds and yet their intelligence had remained consistently at the imbecile level.

Even with much less extended observations a similar extrapolation of previous performance can be used as an effective control. I refer to less extension in time although not necessarily a less extended number of observations. For example, if we give a group of subjects a long series of trials on the pursuit rotor under massed conditions and then insert a short rest interval, performance after the interval will improve remarkably. Had we retained the massing, that is, had there been no rest interval, we would have obtained what we would consider a control score. Obviously, we can't give the same group both conditions so that if we need this control score we would have to run another group to determine how much change took place with and without the rest interval. However, the extended series of trials before the introduction of the rest interval allows us to predict with high accuracy what the score would have been without the rest interval. The 20 (let us assume) massing trials, for which we have a score on each, serve the same purpose as the intelligence tests given each year for 20 years in the above problem. In short, we can determine with very little error what the performance would have been without the rest interval merely by extrapolating beyond the 20 massed trials. Thus, we can use a single condition—the experimental condition—and arrive at the same conclusion at which we would have arrived with the usual control condition. Control measurements are not overlooked in these cases; rather, they are obtained by projecting a stable performance curve to predict what would have happened had the same conditions been retained. Of course, justification for the projection depends somewhat on the stability of the performance curve and background information concerning the phenomenon under investigation. In any case, do not let my fervor to make you sensitive to the need for control groups obscure the realization that there may be a few situations in which the control may not be needed.



the additional control is needed) in that if a significant change in behavior occurs, the specific causal conditions cannot be identified with a specific phenomenon when it was the purpose of the research to so identify the condition and phenomenon. I shall give you a number of illustrations since deficiencies in these matters are fairly widespread.

1. According to a theory (18), one of the processes involved in learning is that the subject must differentiate among stimuli if different responses are to be attached to these stimuli. An implication of the theory is that if there are different degrees of similarity among stimuli, learning rate will be inversely related to the degree of similarity. This implication has been confirmed many times. But, another implication is that if stimuli have some apparent degree of similarity, and subjects are given practice in making discriminations among the stimuli (predifferentiation) before the learning task is instituted in which different responses are to be attached, learning will be facilitated. Several studies (e.g., 17, 27) have supported this expectation. Briefly, predifferentiation consists of presenting the stimuli to the subject in some sort of discrimination learning until he clearly can differentiate one from the other. The subject is then given a new task in which he must attach new responses to each stimulus. The control group which has been used in these experiments is not given the predifferentiation experience; it is given only the final test task. Diagrammatically, the two conditions appear as follows:

	<i>Predifferentiation?</i>	<i>Test Task?</i>
CONTROL:	No	Yes
EXPERIMENTAL:	Yes	Yes

If comparison of performance on the test task shows the experimental group to be superior to the control group, it is attributed to the predifferentiation experience.

If the details of the above procedure are adequately carried out (as they have been in these experiments), there is no denying the conclusion that differences in the test task are due to the predifferentiation experience. But this general conclusion is not the one at which the investigators in these experiments arrived. The conclusions referred to the specific differentiation among stimuli as being the

sign of the transfer. As it stands, I do not know whether in absolute magnitude the performance indicates negative, positive, or negative for some points, positive for others. I *do* know that I have a variable which significantly influences the amount of transfer.

But now let us suppose the results from this experiment were as depicted in the lower half of the figure. I think the first tendency is to say that according to this graph transfer is not related to degree of first-task learning. Of course, we could say that between 5 and 20 trials of first-task learning there is no relationship with transfer. But if we had a control group we might find that all points show negative transfer, positive transfer, or zero transfer. It may be that transfer is related to degree of first-task learning up to about five trials, after which it levels off. In any case, the control group greatly broadens the conclusion which can be reached. (In this particular illustration, by careful planning it is possible to get control measurements by using performance on the first list; but in many such experiments this is not possible.)

So much for this matter; it will be a rare case when one or more control conditions or groups do not add appreciably to the conclusions of the experiment even though a systematic exploration of a stimulus dimension has been undertaken.

#### FAILURE TO USE APPROPRIATE OR NECESSARY CONTROL GROUPS

In research where no control group is used, and where there is essentially only one treatment, the data may demonstrate a significant change in behavior from pretest to the posttest. The ambiguity lies in the fact that it is impossible to tell whether the change resulted from the conditions inserted by the investigator or from some factor or factors which occurred between the two testings. Even if no change is found from the first testing to the next, a conclusion that the variable is ineffective is not completely acceptable because a variable influencing behavior in a contrary way may have operated.

In the cases now to be discussed, the error is not one of failing to use a control group but failure to use an appropriate control or failure to use an additional control group that is necessary to arrive at the conclusion desired. The control group is inappropriate (or

and 24 control patients. The control subjects were even treated to the point of preparing them for surgery and taking them to the operating rooms but, of course, no surgery was performed. Members of the experimental group had various parts of the frontal lobes rendered ineffective by surgical procedure.

Many, many tests were given to both groups before and after the operation and some differences in change from the pretest to posttest were noted between the two groups. Figures on subsequent discharges tended to favor the experimental group. Thus, the results, while not strongly recommending the operative procedure, were not completely negative to it.

The purpose of this experiment on frontal-lobe operation was to determine the effects of cutting of these lobes *per se*. I think it is a fair evaluation to say that this research did not fulfill its purpose because the control-group condition was inadequate. The assault upon the skull or skull cavities, no matter how large or small it may have been for members of the experimental group, should have been reproduced in members of the control group; all conditions should have been exactly the same for the two groups except the surgical contact with the frontal lobes for the experimental group. Only by such a procedure can a conclusion be reached about the influence of frontal-lobe cutting. That this appropriate control procedure was not used is a little perplexing since an effect of various forms of shock on mental illness was being suspected in many hospitals and it would have been reasonable to believe that surgical shock may be a counterpatt of other forms of shock treatment. As a matter of fact, scattered evidence could have been brought together suggesting that if the operative procedure has any appreciable influence, it is due to the shock accompanying the operations and not the cutting of the lobes.

3. We turn next to a quite different field of research and again evaluate a study where certain control procedures were used but in which a critical control is missing (8). The phenomenon dealt with was *assimilation*. By this is meant the tendency for objects to appear like the typical object of the class to which they belong. I shall simplify the conditions of the experiment in the interest of brevity. A figure was flashed on a screen by a tachistoscope and the subject was asked to draw immediately what he saw. The figure might be

factor which produced the facilitation in the performance on the test task for the experimental group. That is, the results were taken to support the theory that the predifferentiation experience reduced the effective similarity among the stimuli. As we now view this situation it is believed that the use of this control group is inadequate to support the conclusion. Again, the discussion must refer back to priority of concepts. The predifferentiation experience given the experimental group may have allowed for the operation of two well-established phenomena either of which might have produced the facilitation on the test task. One of these phenomena is called *learning-how-to-learn* or *practice effects*, and the other, *warm-up*. The predifferentiation experience may have allowed these phenomena to operate so that the transfer to the test task may not be the result at all of the discrimination presumably set up among the stimuli by the experience given the experimental subjects. The appropriate procedure would be to give the control group predifferentiation experience on a task in which the stimuli were different from those used in the test task. Thus, the idea is to allow learning-how-to-learn and warm-up effects to influence the performance on the test task equally for both groups. This leaves only the predifferentiation of stimuli as the difference between the groups so that if the performance of the experimental group is better than that of the control on the test task, it can clearly be allotted to the experience of the subjects with the particular stimuli used on the test task. And we would have given rise to a new phenomenon over and above two already established phenomena. Actually, when appropriate controls have been used there is evidence for the effect of predifferentiation in some situations (e.g., 24) and not in others (e.g., 3). Of course, I think it is apparent that if one wanted to use three control groups it would be possible to determine how much each of the three factors (learning-how-to-learn, warm-up, predifferentiation) is contributing to the performance on the test task.

2. One of the most extraordinary research projects of postwar years attempted to give a conclusive answer to the question of whether or not various forms of frontal-lobe operations helped certain types of mental illness (25). The project involved psychologists, psychiatrists, and surgeons. There were 24 experimental patients

unwarranted for it has not been demonstrated by these procedures. To do this, some sort of a control group is needed in which instructions of "what to expect" are given, but either (a) nothing is flashed on the screen, or (b) a figure entirely irrelevant to the instructions flashed. I cannot be sure which would be the appropriate control without preliminary investigation. Flashing nothing but a flash would be ideal if the subject could be made to believe that something besides just a flash was being flashed. The major point is that the subjects in the experimental group (instructed what they would see) might have drawn exactly what they did draw even though they saw nothing, or even though they saw something quite different from what they were told to expect. This being the case there could have been little or no interaction between the instructions and what they saw on the screen, that is, there could have been little or no perceptual assimilation. The conclusions reached by the investigators are questionable until a control group treated in some such way as suggested is inserted in the design.

4. I suspect that *time* is the most frequently manipulated environmental variable in all of psychological research. Also, it occurs frequently as a confounding variable. I will give four illustrations of the types of errors which are made.

This first design is fictitious. An investigator developed a theory of transfer which predicted that a noxious stimulus given immediately after learning the first task would, to put it crudely, "blot out" the first task so that there would be less transfer than if no noxious stimulus were given. Furthermore, the theory predicted that the longer the "rest" following the noxious stimulus, the less the effect. It was this latter prediction he was interested in and to test it he used the following three groups:

GROUP I: Task A; raucous buzzer for 1 minute; Task B

GROUP II: Task A; raucous buzzer for 1 minute; rest 15 minute; Task B

GROUP III: Task A; raucous buzzer for 1 minute; rest 1 hour; Task B

The theory would thus predict that transfer would be greater for Group III than for Group II, and greater for Group II than for Group I.

It is quite clear that time between Task A and Task B differs for the three conditions. Group I is the nominal control group for

perceived as one of two objects which had rather high similarity. Although several figures were used, I shall use only one as illustration. A figure was used which might be perceived as a lima bean or as a canoe. The subject was first presented this for a very short exposure period, namely, 10 milliseconds. But, after this presentation he was asked to draw what he saw. The figure was then presented for successively longer exposure times, the subject drawing what he saw after each exposure. Finally he was shown the figure for as long as he wished and was asked to draw it as faithfully as he could. He was then asked to draw a figure which would look as much like a lima bean as possible and one which looked as much like a canoe as possible. Judges then rated the drawings made under tachistoscopic presentations by using as reference points the three drawings made without time limit of presentation.

The two critical conditions were:

**CONTROL:** Simply told to draw what they saw following each tachistoscopic exposure.

**EXPERIMENTAL:** Told that they would be shown a canoe (or a lima bean) before the series of increasing length of tachistoscopic procedures was started.

The critical comparisons consisted in whether or not the subjects who were told they would see a canoe (or lima bean) drew a figure more like a canoe (or like a lima bean) than those who were told nothing. This in fact they did. That is, the subjects who were told that they would be shown a lima bean drew figures more like a lima bean than those who were told nothing and their figures were more like a lima bean than those who were told they would be shown a canoe. Differences in the drawings decreased somewhat as exposure time increased but not markedly.

The primary conclusion from the results is that assimilative memory changes which are supposed to take place over long retention intervals also take place in reproductions drawn immediately after perception. The perception is assimilated or modified to represent the typical member of the class to which it belongs when it is presented in an ambiguous fashion (as in the tachistoscopic method). The conclusion assumes an interaction between the instructions of "what to expect" and the figure flashed on the screen. But this is

6. Another study (15) manipulating the time variable had its origin in an aspect of psychoanalytic theory which suggests that trauma is one cause of neurosis. This particular study set out to see if trauma could be reduced in intensity by allowing it to occur in small doses. What this reduces to is that if the subject is exposed to a trauma-producing situation there will be less trauma if exposure is by distributed practice than by massed practice. The subjects were 21 puppies about 16 weeks old. The trauma-producing situation was a box which was so small that the puppy could just barely turn around. Some pilot observations indicated that confinement in the box produced considerable agitation and this could be measured by number of movements the animal made and the number of yelps emitted. Subjects in the massed group received 10 continuous minutes in the box; as far as the time variable is concerned this is the control condition. Those subjects in the distributed group had 1 minute in the box, 1 minute out, 1 minute in, and so on, until a total of 10 minutes had been spent in the box. The results show that the number of yelps made by the massed group was significantly greater than the number made by the distributed group. Number of movements did not differ appreciably.

I am not concerned here whether or not this constitutes any sort of test of psychoanalytic theory. Nor am I concerned with whether or not the box was trauma-producing. This might be questioned because three subjects were dropped from the experiment because they voluntarily entered the box. But, on operational grounds there is justification for calling this a trauma-producing situation. My concern is with other matters. Associated with the distributed condition was the fact that the puppies were handled 20 times, 10 in being put in and 10 in being taken out. This handling alone may have been responsible for the difference in behavior and a control in which the handling occurred with massed practice seems necessary; or, complete avoidance of handling in the experimental group might have been accomplished. Another matter is that measurements on the two groups were not taken at comparable time intervals from the start of the experiment. Thus the distributed group had a total of 20 minutes in the general experimental situation, the massed group only 10. Perhaps adaptation to the situation as a function of time took place. If a massed control group spent the full 20 minutes in

there is no rest after the buzzer. Assume that the results show what was predicted by the theory. Can we attribute the differences to time after buzzer? It doesn't seem so. Perhaps transfer would show the same relationship if the buzzer had not been present in any of the conditions; time may be the effective variable, not time after buzzer. It would seem necessary to use three control groups in which no buzzer was given but in which the time between Task A and Task B correspond to each of the intervals in the three experimental groups.

5. In one experiment the investigators (28) wanted to determine the influence of different lengths of delay of knowledge of results on accuracy of drawing three-inch lines while blindfolded. Three different groups were used, each being given a series of trials. The essential aspects of the procedures were as follows:

- GROUP I: Draw line; given immediate knowledge; rest 10 seconds; draw line, etc.
- GROUP II: Draw line; wait 10 seconds; given knowledge; rest 10 seconds; draw line, etc.
- GROUP III: Draw line; wait 20 seconds; given knowledge; rest 10 seconds; draw line, etc.

Roughly, subjects in Group I drew a line every 10 seconds, those in Group II every 20 seconds, and those in Group III, every 30 seconds. Would the performance have varied as a function of these different intertrial intervals even without differences in delay of knowledge? We might judge that it would not have, but such judgments should not have to be made for appropriate control groups would have eliminated this possible confounding of time by time as a source of concern. Two control groups, having immediate knowledge but with rests of 20 and 30 seconds respectively before drawing the next line would have ruled out differences in time between drawings as a confounding variable or would have shown that it was a variable influencing performance. You might also suggest that equating time between successive drawings for all three groups would solve the problem. In a certain sense it does, but an interpretative problem would remain namely, that with variation in delay after drawing a line there would be an inverse relation between knowledge and drawing of the next line. To which time interval would one attribute differences in performance?



cal conclusions when the situation did not justify such conclusions. When a control group with no color-naming was used, but with a task which would prevent rehearsal, differences in error frequency evaporated.

#### SOME SPECIAL PROBLEMS RELATED TO USE OF CONTROL GROUP

Much of the material to be covered in this section is drawn from three reports, one by Solomon (30), one by Campbell (9), and one by Hovland *et al* (21). I have suggested earlier that good designs yield still better designs as a result of intensive analysis of data and a consideration of their implications. The betterment consists largely in being able to identify more and more specifically the conditions which influence behavior; gross phenomena are broken down into parts and the causal condition for each part identified. Solomon's article is concerned largely with transfer of training, Campbell's with attitude changes, and Hovland's with attitude changes, but the ideas involved may be applicable to wider ranges of behavior. Let us start the analysis by considering a specific experiment.

A study was done to determine the influence of a specific motion picture on attitude toward Jews (26). The essentials of the design were as follows:

	Pretest?	Movie?	Posttest?
CONTROL:	Yes	No	Yes
EXPERIMENTAL:	Yes	Yes	Yes

In accordance with our previous discussion, we would say that differences in the posttest (if present) must be attributed to differences in the treatment, namely, seeing the movie (we shall assume all other matters are handled adequately). Before restricting this conclusion somewhat, let us review possible factors which may enter into differences in scores between the pretest and posttest and for which the control group serves as a control.

1. We know that being tested twice on the same test, or on equivalent forms of a test, may result in a practice effect or an increase in score for some reason. So, therefore, some of the change from pretest to posttest may be a result of this practice. However,

the box, and comparisons of the behavior during the 1st, 3rd, 5th, . . . and 19th minutes were made with the successive 10 minutes of the experimental group the differences in behavior might have disappeared. As the procedure was actually carried out we may conclude that something which was done to the puppies produced the differences in behavior, but we have no basis for concluding that it was the difference in time *per se* between trials which produced the difference.

7. In studies on distributed practice in verbal learning, a problem which has plagued many investigators is how to "fill" the rest intervals so that the subjects won't rehearse the task they are learning. The desire to prevent rehearsal is understandable since it avoids the very type of experimental error I am discussing. Distributed trials are introduced to discover what effect the rest intervals as such have on performance; therefore, it is desirable that any superiority in performance under distribution not be attributed to the extra learning which might occur with rehearsal. One of the tasks commonly used to prevent rehearsal is color-naming. The subject is simply given a board on which patches of paper of various hues are pasted and during the interval he names the colors at a fairly rapid rate. Subjects report that they cannot rehearse while carrying out this task.

When I started a series of studies on distributed practice a few years ago, color-naming was introduced as a standard rest-interval filler. A finding which persistently showed up in these first studies was that subjects serving in distributed conditions made more overt errors in learning the tasks than did those subjects learning under massed conditions; this was true even though learning was actually more rapid under distribution than under massing. I took this difference in error frequency to have considerable importance for it was relevant to existing theories. However, subsequent research (34) showed that this difference in error frequency was due entirely to the naming of colors and not due to some subtle process taking place during the rest intervals as I had been trying to make out. Certainly the difference in error frequency was caused by the distributed conditions but I had been blind to the fact that what was being used in the distributed intervals to prevent rehearsal had been responsible for the greater tendency to make errors. I had been drawing analyti-

as we have the control group, no bias will attend our results as a consequence of such a factor.

Thus we see that with the use of the control group our conclusion is that any difference which occurs in a posttest must be attributed to the experimental treatment. However, as pointed out by Campbell and Solomon, the experimental treatment may be effective only *because* of the pretest. That is, the pretest sensitizes or prejudices the subject to the topic so that the experimental treatment *does* produce changes in behavior. Thus, if we should use the following design, omitting the pretest, no difference in the two groups on the posttest might occur even though they did occur when we had a pretest:

	Pretest?	Treatment?	Posttest?
CONTROL:	No	No	Yes
EXPERIMENTAL:	No	Yes	Yes

In short, by the traditional method of using a pretest we get an interaction between the pretest and the experimental treatment to produce the difference on posttest. If this is the case then the generalization of our results would hold only for a population that had the pretest and it is quite unlikely that anyone except our subjects would have such a test.

Now certainly such behavioral interactions are an important object of study. (Campbell makes some guesses concerning the subject matter areas in which such interactions may and may not be expected. Also it may be noted that the interaction need not only facilitate the posttest scores, for Solomon found a negative effect in a transfer of training study.) When these interactions are to be studied, we need to add another group which did not have the pretest, but which was exposed to the experimental treatment as follows:

	Pretest?	Treatment?	Posttest?
EXPERIMENTAL:	Yes	Yes	Yes
CONTROL-1:	Yes	No	Yes
CONTROL-2:	No	Yes	Yes

Control 2 in this design allows us to factor out the interaction or sensitization effect of the pretest; any change in the experimental group over and above this must be attributed to the experimental

since both groups have both tests, this practice should not bias the results in favor of one group more than the other.

2. A second source of change from first to second testing is extra-experimental experience. Thus, if in the above experiment subjects happened to attend a sociological lecture in which the topic was "race prejudice," and if they attended it between pretest and posttest, the posttest scores might be influenced by this experience. Here again, however, unless the experimenter has evidence that more in one group attended the lecture than in the other, no bias is present although clearly some of the change from pretest to posttest may be accounted for by this. If two groups in any research have differential extra-experimental experiences between pretest and posttest, it is quite clear that we could have very biased results and arrive at quite inappropriate conclusions as far as our experimental manipulations are concerned.

I wish to digress for a moment at this point. I think it is a truism that the longer the interval between the pretest and posttest, the greater the probability that extra-experimental experiences will influence the results. Nevertheless, there may be many cases where the experimenter wants to do long-term studies; that is, studies in which the interval between pretest and posttest may be several months or even years. Furthermore, he may wish to sample the time dimension at various points throughout the long interval. For example, if in the above experiment the attitudes of the experimental group had changed to a greater extent than the control, he might wish to measure the permanence of the change by testing again, say, after six months. In any such studies over time, the control group must be maintained in order to assess the influence of the experimental variable. If the subject matter is such that repeated testings of the same group is inadvisable, then as many experimental and as many control groups as there are time intervals must be used. Extra-experimental experiences become increasingly important as the time interval grows longer.

3. It is possible that changes may take place between pretest and posttest due to intra-organic growth processes of the subject. As Campbell points out, this could be quite true in young children where neuro-muscular growth occurs relatively rapidly. But, as long

learning, whether dealing with nonsense syllables or prose passages, we have no way of putting the units together into a series of tasks and obtaining any guarantee that the tasks will be equivalent in difficulty. The relative difficulty of the tasks can only be determined by an empirical test. If we do not make this empirical test, we have no alternative but to counterbalance the environmental conditions equally over all tasks so that differences in difficulty (if these exist) will not bias the behavior for any one condition. We may have many ratings on verbal units, e.g., affectivity, meaningfulness, familiarity, but when these units are put together in lists, the lists may differ in difficulty even though the ratings on individual units are equated. At least such tasks differ in difficulty frequently enough so that we cannot proceed with research without using some form of counterbalancing. I shall later have more to say about this problem of handling task dimensions in experimental design. Let me turn now to a concrete illustration of a possible task confounding when manipulating an environmental variable.

2. To simplify the problem unmercifully, I will say that the investigators in this particular experiment were interested in the effect of shock on the learning and retention of verbal materials (5). A serial list of 15 nonsense syllables was used; 5 of these were followed by shock each time they were presented, 10 were not. The learning was carried to a performance criterion and it was found that the shocked syllables were learned more rapidly than the nonshocked. Was this difference due to shock *versus* no-shock? All subjects had the same 5 syllables shocked. If these syllables were less difficult than the other 10, the same results would have been found without shock as were found. If the shocked syllables were more difficult than the others, the findings minimize the difference in performance as a function of shock *versus* no-shock. We have no way of knowing what to conclude from this experiment. The issue could have been resolved by using a control group which learned the list without shock or by systematically changing the shocked and nonshocked words from subject to subject so that, all subjects considered, no bias would have occurred.

3. A research problem which exists in widely different fields revolves around subject biases. Features of experimental tasks may provoke responses reflecting these biases and unless these features

treatment. For certain studies a fourth group having only the post-test will be included, but details of this can be found in Campbell's and Solomon's articles.

If one is not interested in the interactive effects of the pretest but is interested solely in the effect of the experimental treatment as a means of generalizing the results to the population from which the samples were drawn, then clearly we should omit the pretest and simply give one group the experimental treatment, the other group not, then measure for differences. The use of the pretest may in many instances be no more than a useless practice that has grown up over the years. If subjects cannot be assigned at random to the groups then certainly we need some means for checking their equivalence and thus the use of the posttest has grown up as fairly standard practice even if it is quite possible to assign at random. Furthermore, if the pretest is used, and if interactions between it and the treatment occurs, then the design which adds the group not having the pretest (in order to "take out" the interactive effects) certainly must rest on some assumptions concerning randomness of assignment to groups.

The whole point of this section, to repeat what I said initially, is to give a brief demonstration of how control groups, perhaps several in a single experiment, may be added for the purpose of pinning down and isolating effects of specific factors within a complex of factors.

#### CONFOUNDING BY TASK VARIABLE WHEN MANIPULATING ENVIRONMENTAL VARIABLE

It is my belief that errors which fit this category are infrequent. At least, there is some reason to believe that it should be true. The reason is that when manipulating an environmental variable the common procedure would be to use the identical task for all conditions. Yet, there are a few problems which do arise in such research and we should sample them.

1. I have previously discussed the problem of the balancing of progressive errors. Such balancing is also necessary to equalize for task differences when manipulating an environmental variable and when each subject is to serve in all conditions. In the area of verbal

learning, whether dealing with nonsense syllables or prose passages, we have no way of putting the units together into a series of tasks and obtaining any guarantee that the tasks will be equivalent in difficulty. The relative difficulty of the tasks can only be determined by an empirical test. If we do not make this empirical test, we have no alternative but to counterbalance the environmental conditions equally over all tasks so that differences in difficulty (if these exist) will not bias the behavior for any one condition. We may have many ratings on verbal units, e.g., affectivity, meaningfulness, familiarity, but when these units are put together in lists, the lists may differ in difficulty even though the ratings on individual units are equated. At least such tasks differ in difficulty frequently enough so that we cannot proceed with research without using some form of counterbalancing. I shall later have more to say about this problem of handling task dimensions in experimental design. Let me turn now to a concrete illustration of a possible task confounding when manipulating an environmental variable.

2. To simplify the problem unmercifully, I will say that the investigators in this particular experiment were interested in the effect of shock on the learning and retention of verbal materials (5). A serial list of 15 nonsense syllables was used; 5 of these were followed by shock each time they were presented, 10 were not. The learning was carried to a performance criterion and it was found that the shocked syllables were learned more rapidly than the non-shocked. Was this difference due to shock *versus* no-shock? All subjects had the same 5 syllables shocked. If these syllables were less difficult than the other 10, the same results would have been found without shock as were found. If the shocked syllables were more difficult than the others, the findings minimize the difference in performance as a function of shock *versus* no-shock. We have no way of knowing what to conclude from this experiment. The issue could have been resolved by using a control group which learned the list without shock or by systematically changing the shocked and nonshocked words from subject to subject so that, all subjects considered, no bias would have occurred.

3. A research problem which exists in widely different fields revolves around subject biases. Features of experimental tasks may provoke responses reflecting these biases and unless these features

are adequately balanced, the biases may influence performance under one condition more than under another. These biases are commonly called constant errors. In psychophysical experiments many of these have been named (e.g., space error, movement error, habituation, and so on) and are, of course, phenomena for study by research in and of themselves. But, unless the investigator is interested in these constant errors elicited by the task presented the subject, he designs his experiment so that they will not influence one condition more than another.

In animal experimentation these constant errors are a continual headache and rather extreme steps must be taken to prevent them from differentially affecting the results for conditions of the experiment. We may think of these in another way. Certain features of a task may have more "cue value" to an animal than others; that is, because of certain experiences or because of genetically determined reasons, all features of a task do not have equal probability of being attended to. If the investigator wants the animal to attend to a particular set of cues, all other cues must be balanced so that they will not bias the results—so the investigator may state precisely the nature of the task presented the animal. Let us take an example; let us suppose that we are going to determine the influence of the effect of magnitude of reward (an environmental variable) in learning a black-white discrimination. To simplify the problem, assume we have two reward magnitudes, small and large, and a different group of rats for each of these two conditions. A jumping stand is used as the apparatus in which to conduct the experiment. Since we want the rats to learn on the basis of black-white discrimination, we have several balancing procedures to accomplish if we want to get an unbiased estimate of the influence of the magnitude of reward on the black-white discrimination problem.

First, our rats may not have equal propensities for black and white initially. Assume that we ignored this and arbitrarily chose to put the food for both groups behind the white card. Assume further that the rats had a strong white-going bias. Our results might show very rapid learning under both conditions and we might conclude that magnitude of reward was not an effective variable. Actually, the animals simply didn't learn anything new; they simply executed a habit to the situation. To avoid such a possi-



bility, we might have half the rats trained on white and half trained on black. Or, we might determine the bias for each rat and then make the positive card (the one which brings reward) opposite to the bias.

Rats may have right- or left-going biases also, so we wouldn't always have the white card on the left and the black on the right. If we kept the card of a given brightness in a constant location the animals may learn on the basis of position cues. Our object is to make sure that the animals cannot learn the correct response based on any cue except the black-white cues. Only if this objective is attained can we state the relationship between magnitude of reward and rate of learning a black-white discrimination. And of course, the most serious confounding would occur if the small reward was always placed behind, say, the black card and the large reward behind the white card. Here we would have a clear-cut instance of task confounding when manipulating an environmental variable.

I have no further illustrations of this type of confounding. As I said earlier, in most experiments where an environmental variable is being manipulated the same task is used for all conditions so that such confoundings are automatically eliminated. Yet, I do not think we will waste our efforts if, before we do an experiment of this nature, we ask ourselves whether the particular task we plan to use will favor one environmental condition over another. There may be more possibility for subtle errors here than I suspect.

#### CONFOUNDING BY ENVIRONMENTAL VARIABLE WHEN MANIPULATING TASK VARIABLES

This confounding refers to cell 4 in the table found on page 91. Except for a few areas of research, this confounding has produced little difficulty (if the number of notes in my file is an adequate index). The more obvious possible confoundings of this type are usually automatically handled by the investigator. Nevertheless, let me quickly run over some of these situations if for no other reason than to get the "feel" for the situation.

1. If we are varying meaningfulness of verbal materials, length of arm in a T-maze, number of alternatives on a multiple-choice examination, size of target in distance estimation, intensity of flash on

are adequately balanced, the biases may influence performance under one condition more than under another. These biases are commonly called constant errors. In psychophysical experiments many of these have been named (e.g., space error, movement error, habituation, and so on) and are, of course, phenomena for study by research in and of themselves. But, unless the investigator is interested in these constant errors elicited by the task presented the subject, he designs his experiment so that they will not influence one condition more than another.

In animal experimentation these constant errors are a continual headache and rather extreme steps must be taken to prevent them from differentially affecting the results for conditions of the experiment. We may think of these in another way. Certain features of a task may have more "cue value" to an animal than others; that is, because of certain experiences or because of genetically determined reasons, all features of a task do not have equal probability of being attended to. If the investigator wants the animal to attend to a particular set of cues, all other cues must be balanced so that they will not bias the results—so the investigator may state precisely the nature of the task presented the animal. Let us take an example; let us suppose that we are going to determine the influence of the effect of magnitude of reward (an environmental variable) in learning a black-white discrimination. To simplify the problem, assume we have two reward magnitudes, small and large, and a different group of rats for each of these two conditions. A jumping stand is used as the apparatus in which to conduct the experiment. Since we want the rats to learn on the basis of black-white discrimination, we have several balancing procedures to accomplish if we want to get an unbiased estimate of the influence of the magnitude of reward on the black-white discrimination problem.

First, our rats may not have equal propensities for black and white initially. Assume that we ignored this and arbitrarily chose to put the food for both groups behind the white card. Assume further that the rats had a strong white-going bias. Our results might show very rapid learning under both conditions and we might conclude that magnitude of reward was not an effective variable. Actually, the animals probably didn't learn anything new; they simply executed a habit they brought to the situation. To avoid such a possi-

two conditions, low and high meaningfulness of material. We give two groups of subjects a constant number of learning trials on each task and measure retention after 24 hours. If we get a difference on our retention measurements for the two conditions can we attribute the difference to differences in meaningfulness? We tend to say of course we can. Assuming all other experimental problems have been properly handled, the only possible cause for the differences in retention measurements was meaningfulness. At one stage in the development of our science I suppose such a finding would have been accepted. But, at the present level of development, and in terms of the problem stated, such a finding would be of little analytical worth. The problem is to determine the influence of meaningfulness on forgetting. By the above procedure we do not know whether that variable influenced learning, forgetting, or both. If meaningfulness influenced learning the differences we measured after 24 hours must be attributed to meaningfulness; but is it because of different degrees of learning before the retention interval or because of differential rates of forgetting due to the intrinsic nature of the material, or both? There is no way to tell from such a set of data. But, knowing as we do that strength of association is a powerful variable determining forgetting, we realize that in order to determine the effect of meaningfulness on forgetting, strength of response must be equivalent for the two levels of meaningfulness before the retention interval is introduced. In short, such an experiment in which a task variable is manipulated is confounded by an environmental variable (degree of learning). We have operationally distinguishable phenomena, learning and forgetting, and our confounding does not allow us to tell which phenomenon is being influenced by the variable. There are several studies in the literature (e.g., 2, 13) using various task variables in which no consideration was given to this distinction between learning and forgetting. Since our variables may influence one and not the other, we must keep our references clear.

To avoid such confoundings it became rather common practice to carry acquisition to a given level of performance for all conditions. Thus, we might take both groups, learning materials of different meaningfulness, to a criterion of one perfect trial on the assumption that by so doing the strength of responses or degree of learning was

alpha waves, and so on, we are manipulating task variables. In order to assess unambiguously the influence of these task dimensions on behavior the environmental variables must be constant for all conditions unless it is known that the specified environmental difference does not influence the behavior being measured. In running a T-maze experiment the number of potential confounding environmental variables is very large. Rats in different groups should be run at the same hour of the day because of diurnal variations; the "smells" around the maze should be constant for all groups; temperature should be the same, and so on. If we are using the psychogalvanic response and determining its relation to varying degrees of affectivity of words even the humidity must be held constant for each condition. In all cases, of course, when I say "held constant" this does not mean that there cannot be any variation (although this would be desirable); it means that if there is variation in these possible confounding variables, the variation is equivalent for all conditions so that no bias can enter.

So much for such routine matters. I want to turn now to a rather subtle confounding which arises in certain research areas.

2. So far as I can determine, the confounding about which I wish to speak now would occur only in certain types of learning experiments. More specifically, these experiments have two stages, namely, an acquisition stage and then some subsequent test for a different phenomenon. These would include studies of retention, extinction, and transfer. There could be three stages involved, as in acquisition, extinction, and spontaneous recovery, and these must necessarily be studied in that order. When we have these two-stage (or more) experiments and we wish to determine the influence of a task variable on the second-stage phenomenon, we must be sure that the performance at the end of the first stage was equivalent for all conditions. The source of the design problem is the fact that if the task variable influences first-stage acquisition it is very difficult to be sure that the conditions are equal on one very relevant variable, namely, strength of association, at the end of the first stage. Let me translate this into a concrete illustration.

Assume that we wanted to determine the influence of meaningfulness of material on rate of forgetting. In order to study forgetting (second stage) we must first have learning. Suppose that we have

is not allowed to fluctuate because it is known that at some frequencies differences in intensity will produce small changes in phenomenal pitch. The same would be true in the visual modality. If luminance is being varied to determine its influence on some visual phenomenon, hue or wave length is held constant. Two investigators (22) had the idea that judgment of depth would be related to the brightness of colors of the objects whose depth was being judged. To work this out in the laboratory the traditional depth-perception apparatus was used in which there are two rods or tubes. One of these is fixed, the other variable. The subject's task is to adjust the variable rod so that it appears to be the same distance from him as is the fixed rod. In this experiment, the fixed rod was always gray, the variable was one of six colors varying in brightness. The subjects were each given 100 trials with the gray tube always on the left. While this might introduce a bit of a space error I am not concerned with this. The results show that brighter colors are judged nearer to the subject than they actually are and the darker colors are judged further away than they actually are. But it seems to me that brightness and hue are confounded in this experiment. Are the differences due to brightness differences, to hue differences, or to both. We cannot tell; to do so would require variation in brightness with hue constant.

2. In Chapter 2, when discussing problems in dimensionalizing task characteristics, I pointed out the very serious problem of obtaining unitary dimensions. Since the reduction of task characteristics to unitary dimensions has not been accomplished with most verbal materials used in learning experiments we must be continually aware of the possibility that some unknown (or at least undimensionalized) characteristic may be partially correlated with the dimension we wish to manipulate. The problem, in miniature form, is much like the one discussed at the end of the previous chapter; that is, the problem of correlated subject variables. But even if we have dimensions of units of verbal material which are unitary and exhaustive, the problem is not completely handled because we must place these units together in a task or list; these problems were discussed earlier. Let us see how this could disturb a simple study on transfer.

Let us suppose we are going to use verbal tasks in a simple study of transfer from one list to another as a function of the similarity

equivalent before the retention interval was introduced. Then, after 24 hours, differences which occur would be attributed to differential rates of forgetting of the material. Certainly we would agree that this design comes nearer to equalizing the degree of learning than does the previous method where number of trials was constant. However, even this is unsatisfactory for precise analysis. The reason is that if the manipulated task variable produces differences in acquisition rates, the two groups do not have an equal degree of learning as a result of attaining the same criterion because of the different rates in reaching the criterion. If we measure the degree of learning on the trial immediately following the criterial trial we would find they were not equal; thus, retention measurements might still be confounded by degree of learning. The magnitude of the confounding will be directly related to magnitude of differences in rate of acquisition. There are at least two possible solutions to this problem; these have been given elsewhere (36, 38) so I will not repeat them here. I will say only that in the case of retention studies an adequate equation of degree of learning before the retention interval has resulted in considerable change in our beliefs concerning the influence of certain task and subject variables on retention.

#### CONFOUNDING BY TASK VARIABLES WHEN MANIPULATING TASK VARIABLE

This refers to cell 5 and represents a confounding which I have judged not only to occur with some frequency in research but also to be one for which we have no pat solutions. It will be remembered that task dimensions may be measured along physical scales or psychological scales. Although task confoundings may occur when manipulating a task variable along a physical scale this seems to be less likely (and the solution easier) than when the dimension is a psychological one. I will give illustrations from both areas but my major emphasis shall be on psychological task dimensions. However, let us start with the case of a task dimension measured along a physical scale.

1. If one wishes to determine the effect of variation in cycles per second of the sound wave on phenomenal pitch, all other dimensions of the sound wave are held constant. Intensity, for example,

the differences in difficulty are fairly great. If the differences in difficulty are large, we had better try a new experiment.

3. I am responsible for one of the most beautiful illustrations that can be found of *task confounding when manipulating a task variable*. I have analyzed this situation elsewhere (35) so shall mention it briefly here. This confounding was between intralist similarity and interlist similarity. I was interested in interactions between intralist similarity and distributed practice on learning and retention. But, at least when using nonsense syllables for the learning material and the same subjects in all conditions, manipulation of intralist similarity led to an inverse "manipulation" of interlist similarity and a consequent *distortion of retention measurements*. This cannot happen when subjects are used in only one condition. Needless to say we have dropped counterbalanced designs for studies of this type.

4. Here is an experiment dealing with the retention of different materials (32). The problem was to study recognition of three different materials judged to have different values for invoking ego-involvement. Each subject was given three cards. On one card the subject wrote his given or first name. On a second card he copied a slogan, the same slogan being copied by all subjects. Finally, on the third card each subject was asked to copy a one-inch square. The subjects were divided into five groups with retention tests being given after a different interval of time for each group. On the retention test, using the recognition method, the subject was given the three stacks of cards resulting from putting cards of all subjects together. From each stack the subject was asked to choose his own particular card. The results were presented in terms of correct recognition of names, slogans, and squares after varying intervals of time.

The idea of the experiment was that the person's own name would invoke the greatest ego-involvement, the slogan less ego-involvement, and the square still less, and the recognition would be directly related to this ego-involvement. We might object to bringing ego-involvement into this picture at all but it is not with that point which I am concerned. Take the two extreme materials, names and squares. Obviously all the subjects would not have the same name. The investigators, therefore, included at least four cards for each name so that recognition would not occur solely on the basis of the

between the lists. We will use 10-item lists made up of units which have been carefully scaled for similarity so that we can clearly differentiate, say, three degrees of similarity. To simplify the procedure we will further assume that we use three random groups of subjects, one group for each degree of similarity. All groups learn the same list for their first list. We have constructed three second lists, one which has high similarity with items in the first, one with medium similarity, and one with low similarity. Thus, the items in the three second lists are different. Now suppose we conduct the experiment with our basic measure of transfer effects being some performance on the second list. If we get differences can we attribute these to the similarity variable? Although we have a number of published studies which have used this procedure, it seems to me that we can draw no conclusions concerning transfer as a function of similarity at this point. Since the three second lists were different lists they may have varied on characteristics which influenced their difficulty so that the differences observed may have been a function of this difficulty rather than a function of the similarity between the two lists. The second lists might have differed in meaningfulness, in intralist similarity, in affective tone, or perhaps other factors.

How do we solve this problem for the transfer experiment? Fortunately it can be solved either empirically or by appropriate design. What we need to do is discover if the three second lists do or do not differ in difficulty when not preceded by the first list. We might do this by having three control groups which learn only a second list. It is somewhat more convenient to handle by having half of the subjects in each group reverse the order in which the two lists are learned. Indeed, in this particular case we could have all subjects learn in reverse order and then determine transfer effects on the single common list. This latter procedure would be satisfactory if we do not find appreciable difference in difficulty for our three lists but if we do, we are then faced with possibilities of differences in degree of first-task learning. However we do it, we must show that the lists do not differ appreciably in difficulty; if they do we must, of course, make adjustments in our estimates of the transfer effects which can be attributed to differences in similarity. There are several ways by which this can be done but none is satisfactory if



of learning. We got wide differences in error frequency and a small but insignificant difference in learning which was in favor of the guessing group. A perceptive reviewer pointed out that even this small difference might be attributed to the fact that when guessing the subject may have hit upon a correct response and, having hit it accidentally, may have fixated it. Had this effect been grossly amplified so that differences in learning were significant, we might well have attributed the differences to something intrinsic in the process of making errors and not to the fact that in making more errors the subject had a higher probability of hitting upon the correct response.

5. I have no further specific illustrations to give, but I can imagine a number of situations in which the confounding of task variables by other task variables could easily occur. Supposing we wanted to determine the influence of similarity of multiple-choice alternatives of a paper-and-pencil test on scores on the test. Could we vary similarity among the alternatives and keep meaningfulness, relevance, *et al* of alternatives comparable while varying similarity? Could we manipulate *threat-provoking capacity* of prose passages and have those passages equal on all other factors which might affect performance? If we vary the political slant of speeches to determine effect on learning can we keep all other dimensions of such material equivalent? Whenever we manipulate a task variable based on a psychological dimension we are confronted by a potentially dangerous research situation. A careful study of other possible ways by which the material might vary in addition to the way we want it to vary will prevent us in many cases from arriving at questionable cause-effect conclusions.

#### CONFOUNDING BY ENVIRONMENTAL AND TASK VARIABLES WHEN MANIPULATING SUBJECT VARIABLES

I am placing cells 7 and 8 together in this final section for there is little to say which has not already been said in the previous sections. The major research problem in manipulating subject variables centers around confoundings with other subject variables and this situation has been discussed at length. However, confounding by task and environmental variables may take place in exactly the same

name. Nevertheless, this is a quite unsatisfactory expedient, simply on a probability basis. Assume there were 100 subjects. With four of each name the probabilities of choosing the correct one by chance would be one in four, whereas the probabilities of choosing the square by chance would be one in 100. Add to this the differences which may have existed in color of ink, peculiarities of writing, and so on which would serve as cues for recognition over and above any idea of ego-involvement in one's writing and we see that the results of the experiment just do not seem to have any bearing on the very real problem of motivation and retention.

At this point I would like to add a general caution concerning this matter of biasing results by conditions which do not have equal "guessing potential." Suppose we are manipulating a task variable (or environmental variable or even a subject variable) and the particular variations allow for different response probabilities based on guessing. But, if differing guessing potentials is not the reason for manipulating the task variable, we have a confounding. I mentioned this matter when considering a study in the previous chapter but its full implication needs to be seriously considered in almost every experiment. We must remember that guesses are seldom random; they usually reflect response biases and if our manipulations allow for biases of different strengths to operate we may err in our interpretation. Thus, if we do a study on verbal threshold recognition (e.g., 31) as a function of frequency of usage of words or letters, and if the subjects are instructed to guess, their guesses are likely to be those letters or words with greatest frequency of usage. The subject may not actually see the high frequency words or letters any sooner than those of low frequency but if he guesses he is most likely to guess those of high frequency and thus appear to have seen them sooner.

This guessing may come up in many situations and all we can really do to protect ourselves is to analyze carefully the situation to see if guesses can be made and to ask whether they will influence the conditions differentially. A student of mine once did a study (29) in verbal learning in which one group of subjects was instructed to guess frequently and another was instructed never to guess. Our interest was in producing wide differences in overt errors to see if in turn there was any relationship between these frequencies and rate

because of a special research problem which is, in a sense, a combined statistical-design problem which I want to mention briefly at this point.

In my discussion of confounding of stimulus variables. I intentionally kept the designs simple. For the most part these confoundings are in no way eliminated or mitigated by research in which two or more variables are simultaneously manipulated. That is, the reasoning we have applied to the simpler one-variable experiment may also be applied to each variable in a multivariate design. However, as empirical and theoretical analysis of an area of research develops, multivariate designs may become almost mandatory if adequate statistical tests of certain phenomena are to be made. For example, if one has an hypothesis that drives summate the design for testing this can be quite simple. But, if the hypothesis specifies an interaction say, between strength of drives and the summation function, an orthogonal design (in which both variables are manipulated simultaneously) is virtually necessary in order to make a statistical test of the interaction phenomena. In recent years it has become necessary to make distinctions between variables which influence associative processes and those which influence only performance. To give this distinction empirical substance it is often necessary to use orthogonal designs; at least, such designs are a very efficient way to provide the separation. For example, in varying intensity of stimuli (conditioned or unconditioned) in conditioning experiments the variable may be orthogonal to itself in successive stages of performance as a means of separating the two components (e.g., 19).

Occasionally a variable is by its intrinsic nature constituted of two or more components, either of which may influence behavior. In order to determine how much each is contributing to performance the multivariate design again provides a most efficient way. To illustrate this, consider the variable of ratio of reinforcement in learning studies. This refers to ratio between number of trials given and number of trials reinforced. Thus, we might give one group 100 trials with reinforcement after each trial and another 100 trials with reinforcement after every other trial (on the average). Behavior might differ in two such conditions either because of total reinforcement received or because of something intrinsic to the pattern of

way as they do in cells 2 and 4. That is, when manipulating a subject variable, confounding by a task variable may take place in the same manner as confounding by a task variable when manipulating an environmental variable (cell 2). And, confounding of a task variable by an environmental variable (cell 4) is analogous to confounding by environmental variable when manipulating a subject variable. As was the case for cell 4, these confoundings in cell 7 usually occur when there is a two-stage experiment. Such confoundings may have occurred (e.g., 2, 13) but I see no point in reporting these since their solution requires no special techniques not already discussed.

### RESPONSE ANALYSIS

The discussion of research errors thus far has been concerned with stimulus confounding. I want to turn now to research problems which may loosely be said to arise in analyzing and interpreting response measurements. Again, "errors" in this aspect of research may be minor or major depending upon the magnitude of the discrepancy between what is concluded and what can properly be concluded.

### STATISTICAL ANALYSIS

None of us needs be reminded that erroneous conclusions concerning behavioral relationships are drawn as a consequence of inadequate or inappropriate statistical analysis of response measurements. Graduate training programs for research workers in psychology are including an increasingly heavy amount of formal course work in statistical and mathematical techniques. Psychologists have not only been remarkably ingenious in devising statistical procedures to handle data-analysis problems peculiar to our science but also have avidly seized upon techniques developed in other disciplines and fitted them to our problems. One need only examine research reports published 20 years ago, indeed, 10 years ago, and compare these reports with current ones to perceive the amazing changes which have occurred in the statistical handling of data. As previously indicated, I chose to omit statistical errors from the present survey. I mention them now in part for sake of completeness and in part

inappropriate since such differences in frequencies would be expected under a single condition at different stages of learning.

2. Assume that we are interested in social interaction of small groups when the individuals brought together in these small groups are strangers. Furthermore, assume we hold the hypothesis that the greater the strain or hard work required of the group the higher the morale that develops. To test this we give three groups problems to solve but with the problems for one group being relatively easy, those for another somewhat more difficult, and those for the third, very difficult. Under appropriate incentive conditions the groups work until the problems given them are solved. We measure how long it takes for each group to solve its problems and, as a measure of morale, the number of times the word "we" is used during the problem-solving session. We find that time to solve increases with difficulty (as expected) and that total number of "we's" likewise increases. This might seem to support our prediction but we can see that the longer the time to solve the problem the longer the period during which "we's" can be spoken. The response measure should be converted into, say, number of "we's" spoken per unit of time worked.

3. In a study of problem-solving (1) experimental and control groups were ostensibly given the same problems to solve but by the nature of the variable manipulated, the experimental subjects would necessarily have to take longer to show that they could solve the problem. Yet time-to-solve was presented as a critical response measure.

4. Assume that we introduced a new micrometer into a factory inspection system. To discover its effectiveness we have one group continue using the old system and another the new one. After a month, we count number of rejects produced by the two systems and find there is no difference. But unless the number of rejects is considered in conjunction with total number of products inspected our response measure does not have the meaning we want it to have.

You may feel that the above illustrations represent errors so obvious that you would never make such a mistake. I hope not, but it is at least worth a warning to carefully inspect your response measure to see if it can possibly be an artifact or psychologically

reinforcement. In order to filter out the influence of each factor separately we might use a three-variable design in which magnitude of reinforcement, number of trials, and ratio of reinforcement were all manipulated orthogonally to each other.

#### INAPPROPRIATE RESPONSE MEASURES

We sometimes think that certain responses are inappropriate because we don't believe they are relevant to the behavior being studied. I would not care to take up such criticism on strictly scientific grounds but I might on social-scientific grounds. For the linking of a response measure with a phenomenon is essentially a matter of definition and if our response measure is reliable, it is hard to quarrel with such definitions on scientific grounds provided there is no transgression on already defined phenomena. But, with extreme cases, quarrels can easily be instigated. If I defined a psychotic as one who can run 100 yards in less than 9 seconds, I feel confident that many would believe I was a very fast runner. So then, the issue with which I wish to deal is not a definitional one for I have covered this earlier; rather, it is a matter of inappropriate response measures in the sense that differences (or lack of them) in the data from an experiment may be a consequence of artifacts in the measuring process. Some illustrations will show the kind of thing about which I am thinking.

1. A study (16) was done to compare rate of learning of similar items when groups of such items were bunched together in the list as compared with the case where they were scattered throughout the list. The two lists were presented for 14 trials. As it turned out, the list with bunched items was learned significantly faster than the list with no bunching. Number of overt errors made was presented as an auxiliary measure and the results show more errors were made in the early part of learning in the bunched-item list but fewer on later trials. These error differences are used to support a theoretical interpretation of the differences in learning rate. However, if the learning scores are adjusted to be equal (irrespective of trials), the error frequencies are likewise equal. In short, comparing error frequencies for the two conditions at different stages of learning is

problems that we cannot handle at our present level of mensuration development.

The second issue is one that confronts us many times but it is only in cases where it is markedly exaggerated that we pay much attention to it. The exaggerated picture can be obtained by examining data collected by members of my undergraduate course in experimental psychology. A card is prepared on which five words are repeated over and over again in random order. The five words are red, blue, green, brown, purple. These words are printed in color but never in the color indicated by the word. Thus, the word *red* is printed in blue, brown, green, and purple, but never in red. As one task, subjects read the words on the card as fast as they can, the response measure being time taken to read. As the second task the subjects name the color in which each word is printed. As might be expected, the second task produces very heavy interference and the time taken to name all the colors on the card we used was about 150 seconds as compared to 50 seconds required when merely reading the words. For each task successive trials (once through the card) were given under massed and under distributed practice. For both tasks a slight increase in time to read under massed trials occurred over a series of trials. Under the distributed conditions performance on both tasks improved. In the case of the highly interfering task the improvement was roughly from 150 seconds to 100 seconds over 8 trials. For the reading task the improvement was from 50 to 45 seconds. Superficially it would appear that distributed practice facilitated the interfering task more than it did the reading task and by any conventional statistical test this would be true. But in terms of significance of behavioral changes is this true? Behaviorally, the reading performance, being so close to the asymptote or a physiological limit could be improved but little and it might well be that the 5 seconds improvement in score actually represents a far greater behavioral change than does 50 seconds in the interfering task. What I am saying, of course, is that our response measure, mean change in seconds, may be reflecting horses in one case and apples in another. Again, I cannot offer a general solution to such problems; but, we must not ignore them.

meaningless. Some very competent investigators have made these so-called "obvious" errors. Let me give you one more fictitious illustration by way of warning.

5. Although we are often belabored about our preoccupation with differences in mean performance in our data, most of us still view differences in variances among groups treated differently as a diabolical caprice which forces us to take time out to adjust these differences so we may proceed with statistical analyses. I suspect our reluctance to make anything of behavioral significance out of differences in variance arises from our feeling that such differences just inevitably occur when we have fairly large differences in mean performance. Yet, we should not overlook the possibility that in certain forms of research the difference in variance may be very meaningful behaviorally with or without mean differences. Whenever we see the possibility that our manipulated variable might cause some subjects to "move" in one direction and some to move in the other, I would say that the differences in variance would have psychological meaning. Thus, if we were doing an experiment on prestige suggestion, we might give subjects prose passages and assign names of well-known authors to them in some cases and not in others. The effect of the names might be to raise ratings of liking-of-passages by subjects who liked previous work of the authors and to lower for those who did not. In such a case our means might reflect no difference between well-known and unknowns but the variances would.

#### NONEQUIVALENT RESPONSE MEASURES

There are two issues involved here, but they converge. The first I shall dispose of quickly. We still occasionally find ourselves comparing changes in apples and changes in horses. For example, it *would* be desirable to know how the forgetting of a pursuit-rotor habit compares with forgetting of a verbal habit but I know of no case in which such a comparison or a similar one has been made which will stand inspection. This matter has been discussed elsewhere (33). I do not believe there is a general solution to this problem. We must face the fact again that there are certain research



scientific productivity. Essentially, I do not think the problem of generalization here is any different from what it is in the staid laboratory study where, say, a group of volunteer Freshmen women are used as subjects. Yet, at the same time it seems to me that the temptation to generalize is greater in the case of the mail-type study than in the case in the female-type study. My point is that we must recognize the limitations to generalization in both situations.

But we do have published mail-survey studies in which the intent is to generalize to a population and here we run into very real dangers. For example, in order to develop new scoring keys for an interest inventory 650 inventories were sent to 650 people holding industrial relations positions (23). Approximately 60 per cent were returned. What can we make of the interest patterns of these 60 per cent that we also know applies to the total group and that allows us to develop a key for predicting success in industrial-relations positions?

#### LABORATORY VARIABLES

I have divided the manipulable variables into three classes, namely, task, environmental, and subject. The most general law we could have, say, of relating an environmental variable and behavior, would be one which maintained its integrity irrespective of the task used, irrespective of the values of other environmental variables, and irrespective of the sample of subjects. Right off hand I don't think of any behavioral law that we have which fits these requirements, nor do I believe we will find such. What are the problems involved in generalizing from a set of data?

*Subject variables.* The matter of generalizing to a population from a sample as discussed above refers, of course, to subject variables. As is well known, the bulk of the relationships derived from laboratory work in psychology are literally applicable only to college students or to white rats. To be very accurate we would even have to say that these laws may not be applicable to college students in general since there has been no systematic sampling. And white rats of certain strains have been favored over others.

As far as I am concerned it is a perfectly legitimate enterprise to study white rats *per se* without any intent of generalizing beyond

## GENERALIZATION OF FINDINGS

With or without the aid of theory, the long-term purpose of research is to develop general laws or relationships. These laws subsume the particular; they envelop the detailed findings. Such laws might seem to be the inevitable outcome of science but this is not quite the case. Rather, degree of generality of laws is determined by research directed by judicious consideration of sampling problems. I will discuss three somewhat different aspects of these problems. Again, if errors occur, they are identified by the discrepancy between what is concluded and what can be concluded.

## SURVEY STUDIES

To obtain a description of attitudes, beliefs, habits, preferences, and so on of a specified population, the entire population or a sample representative of it must be measured. These sampling techniques have been developed to a very high level by organizations of pollsters. It is not my intent to examine these techniques, but I do wish to mention briefly the implications of survey studies which use mail questionnaires since such studies persistently crop up in psychological journals. The major focus is on the nature of the conclusions which can be drawn.

In one study (37) 467 questionnaires or information sheets were mailed to members of the staff of a technological institute. The questions asked each recipient to indicate how many scientific publications he had produced, the number of technical journals regularly read, and a lot of other embarrassing questions. The total number of questionnaires returned was 194, which is 42 per cent. The intent of the questionnaire was to discover what factors (e.g., age, training, work habits) correlate with scientific productivity. Now certainly there is nothing wrong with determining correlations among the various indices; we simply must be very careful in our conclusion concerning the meaning of these relationships. We have no idea as to whether these relationships would obtain among those who did not return the questionnaire. We have two populations, one whose members did return the sheets and one whose members did not. These two populations may well differ on factors related to

these classes were held constant at different values the obtained relationship might disappear, might be modified, or might not change at all.

The number of potentially relevant environmental variables is enormous; the number of actually relevant ones may be small. The solution to these problems may come about in two ways. First, by haphazard differences in environmental conditions from experiment to experiment and from laboratory to laboratory, a wide variety of settings of these environmental variables will have occurred. Thus, if a large number of studies on the influence of a given variable on conditioning has been done, and if other environmental variables have been held constant in each experiment but at different levels, and finally, if the same relationship has been found consistently, a *post-hoc* accounting of these variables attests to their irrelevancy. Secondly, we will have systematic attempts to determine the generality of a phenomenon by determining how it is influenced by other potential environmental variables. The more variables we can change with the phenomenon remaining unchanged the greater the generality of this phenomenon. And of course, each new positive test adds to our confidence that extensions to other situations are likely to give the same results. But in any event, there is no easy solution to this problem just as there is no quick resolution to the problem in the case of subject variables and task variables.

*Task variables.* In certain respects, the path of our science toward attaining generalizations across tasks is more obscure than that for subject and environmental variables. To take a simple illustration, suppose I set about to determine the influence of distributed practice on learning, and I want to be able to generalize across tasks. I could pick a number of different tasks which I think are different (i.e., which involve different or uncorrelated subject skills) but my judgment of what constitutes different skills may be hopelessly inadequate or erroneous as also may be the pooled judgments of many persons. The number of ostensibly different tasks which might be devised is almost without limit. As I indicated in an earlier chapter, I have suggested at various times to my colleagues that we might make a start toward the solution of this problem by assembling a wide variety of tasks which we thought were different and then determining the communalities of the skills by factor analysis. The

white rats. So also can we study monkeys, stentors, atoms, butterflies, ocean waves and mountain tops. But if we say that laws developed from white rats are applicable to college students without empirically testing this we are making a frightful leap. Even if we insist that there *must* be some behavioral laws which hold for all living creatures the correctness or incorrectness of our insistence can only be gauged by empirical studies. Using the white rats or monkeys as a source of hypotheses about human behavior is common practice (and it can work the other way too), but no scientific process that I know abrogates the making of empirical tests of the hypotheses.

It is perhaps too early in the development of our science to be overly concerned about the matter of systematically determining the generality of laws from one species to the next. Certainly within the species *homo sapiens* we have as yet no methodical plan for exploring the limits or generalities of relationships. We are still so engrossed in determining reliable phenomena and variables which relate to them within very restricted populations that any generality we may have has occurred as a result of haphazard or fortuitous use of samples from different populations. So what can we say about this matter? First, we must realize that our present laws may be very restricted because of the restricted range of subjects used. Second, our science must eventually make systematic attempts to determine the generality of laws, not only within species, but across species.

*Environmental variables.* To make a general statement concerning the effect of a given environmental variable, the prerequisite is an adequate exploration along the "length" of that variable. If we examine the influence of intensity of conditioned stimulus on rate of conditioning and explore the range, say, between 70 and 100 decibels, we have inadequately explored the dimension and are in no position to make a generalized statement concerning the influence of this variable even for the restricted set of conditions of the experiment. This is the first issue as I see it for this class of variables. The second is more difficult to bring into perspective. Having determined the influence of a given environmental variable on behavior in a particular situation, we must realize that this relationship holds only for the values at which other environmental variables, subject variables, and task variables were held constant. If the variables within

But, one may persist, since the principles of behavior obtained in the laboratory may be restricted to situations in which certain other variables are held constant, how can we be sure that the manipulated variable which is effective in the laboratory will be effective in the field situation? If distributed practice facilitates the learning of nonsense syllables in the laboratory can I be sure that distributed practice will facilitate the memorizing of a foreign language vocabulary in the classroom? Of course not. The issue is simply another manifestation of the problem of arriving at scientific generalizations about phenomena which are not intrinsic to a highly specific set of conditions. Laboratory studies give ideas about variables which might affect performance in real-life situations, but the degree of confidence with which the generalization can be made from laboratory to field depends upon the degree to which many possible counteracting and interacting variables have been studied. The proof of generality lies in a test for it; the laboratory situation in this context may be thought of as a highly efficient means for identifying variables which may be important for the field, but the proof of this lies in the field test where conditions would be less highly if at all controlled. Not even the most advanced sciences can avoid this test even though the laboratory be constructed to simulate as nearly as possible the field conditions. The laboratory is the home of science, not of technology.

## REFERENCES

1. ADAMSON, R. E. Functional fixedness as related to problem solving: A repetition of three experiments. *J. exp. Psychol.*, 1952, 44, 288-295.
2. ALPER, T. G., & KORCHIN, S. J. Memory for socially relevant material. *J. abnorm. soc. Psychol.*, 1952, 47, 25-37.
3. ARNOULT, M. D. Transfer of predifferentiation training in simple and multiple shape discrimination. *J. exp. Psychol.*, 1953, 45, 401-409.
4. BARKER, R. G., DEMBO, T., & LEWIN, K. Frustration and regression. An experiment with young children. *Univ. Ia. Stud. Child Welf.*, 1941, 18, No. 386.
5. BELMONT, L., & BIRCH, H. G. Re-individualizing the repression hypothesis. *J. abnorm. soc. Psychol.*, 1951, 46, 226-235.

factors would define the general subject capacities required by the tasks and research could then be undertaken on the influence of variables on single tasks which represent each factor. In the area of motor learning some success in this direction has been achieved (14). But I am told that present methods of factor analysis are essentially inadequate for such ventures. If this be so, then I think we should develop methods of analysis which are appropriate. In the long run of our science some systematic analyses of tasks must be accomplished.

As I look back over the problems of scientific generalization as I have discussed them I would hope my analysis is in error and that the task is not as gigantic as it appears to me to be. If one measures his own lifetime research efforts against the known work which appears to lie ahead it measures but as a shovel of sand in the vastness of the Sahara. The saving grace is that even a shovel of grains of sand must be in some way representative of all sand and that the grains within the shovel have an undeniable fascination in themselves.

#### LABORATORY TO FIELD

Experimental research in psychology (as well as in any science), research in which the investigator controls variables, is an abstraction in the sense that it never duplicates a real-life situation. The very fact that variables are controlled makes this apparent. So how, one may ask, can we apply the principles derived in the laboratory to real-life situations? How can we generalize from the laboratory to the field? Why not do the research in the field in the first place?

Let us first understand that the laboratory is not as divorced from reality as some would have us believe. Does the subject leave his hates, his skills, his capacity to learn in the dormitory when he comes to the laboratory? Does he come to a room in which there is no temperature, no stimulation, no social interaction? Is the learning of a list of nonsense syllables or the judgment of distances totally unrelated to what the subject does in real life? Of course not. The only major difference between the laboratory and everyday life is that variables other than the ones in which the investigator is interested are not allowed to "roam" at will.

24. KURTZ, K. H. Discrimination of complex stimuli: The relationship of training and test stimuli in transfer of discrimination. *J. exp. Psychol.*, 1955, 50, 283-292.
25. METTLER, F. A. (Ed.) *Selective partial ablation of the frontal cortex*. New York: Harper, 1949.
26. ROSEN, I. C. The effect of the motion picture "Gentleman's Agreement" on attitudes toward Jews. *J. Psychol.*, 1948, 26, 525-537.
27. ROSSMAN, I. L., & GOSS, A. E. The acquired distinctiveness of cues: The role of discriminative verbal responses in facilitating the acquisition of discriminative motor responses. *J. exp. Psychol.*, 1951, 42, 173-182.
28. SALTZMAN, I. J., KANFER, F. H., & GREENSPOON, J. Delay of reward and human motor learning. *Psychol. Rep.*, 1955, 1, 139-142.
29. SCHEIBLE, H. S., & UNDERWOOD, B. J. The role of overt errors in serial rote learning. *J. exp. Psychol.*, 1954, 47, 160-162.
30. SOLOMON, R. L. An extension of control group design. *Psychol. Bull.*, 1949, 46, 137-150.
31. SOLOMON, R. L., & POSTMAN, L. Frequency of usage as a determinant of recognition thresholds for words. *J. exp. Psychol.*, 1952, 43, 195-201.
32. TRESSELT, M. E., & LEVY, B. Recognition for ego-involved materials. *J. Psychol.*, 1949, 27, 73-78.
33. UNDERWOOD, B. J. *Experimental psychology*. New York: Appleton-Century-Crofts, 1949.
34. UNDERWOOD, B. J. Studies of distributed practice: VI. The influence of rest-interval activity in serial learning. *J. exp. Psychol.*, 1952, 43, 329-340.
35. UNDERWOOD, B. J. Intralist similarity in verbal learning and retention. *Psychol. Rev.*, 1954, 61, 160-166.
36. UNDERWOOD, B. J. Speed of learning and amount retained: A consideration of methodology. *Psychol. Bull.*, 1954, 51, 276-282.
37. VAN ZELST, R. H., & KERR, W. A. Some correlates of technical productivity. *J. abnorm. soc. Psychol.*, 1951, 46, 485-495.
38. YOUNG, R. K. Retroactive and proactive effects under varying conditions of response similarity. *J. exp. Psychol.*, 1955, 50, 113-119.

6. BORING, E. G. The nature and history of experimental control. *Amer. J. Psychol.*, 1954, 67, 573-589.
7. BROWER, D., & OPPENHEIM, S. The effects of electroshock therapy on mental functions as revealed by psychological tests. *J. gen. Psychol.*, 1951, 45, 171-188.
8. BRUNER, J. S., BUSIEK, R. D., & MINTURN, A. L. Assimilation in the immediate reproduction of visually perceived figures. *J. exp. Psychol.*, 1952, 44, 151-155.
9. CAMPBELL, D. T. Designs for social science experiments. Unpublished MS.
10. CHILD, I. I., & WATERHOUSE, I. K. Frustration and the quality of performance. I. A critique of the Barker, Dembo, Lewin experiment. *Psychol. Rev.*, 1952, 59, 351-362.
11. COWEN, E. L., & COMBS, A. W. Follow-up study of 32 cases treated by non-directive psychotherapy. *J. abnorm. soc. Psychol.*, 1950, 45, 232-258.
12. CROWN, S. An experimental study of psychological changes following prefrontal lobotomy. *J. gen. Psychol.*, 1952, 47, 3-41.
13. EDWARDS, A. L. Political frames of reference as a factor influencing recognition. *J. abnorm. soc. Psychol.*, 1941, 36, 34-50.
14. FLEISHMAN, E. A. Dimensional analysis of psychomotor abilities. *J. exp. Psychol.*, 1954, 48, 437-454.
15. FREDERICSON, E. Distributed versus massed practice in a traumatic situation. *J. abnorm. soc. Psychol.*, 1950, 45, 259-266.
16. GAGNE, R. M. The effect of sequence of presentation of similar items on the learning of paired-associates. *J. exp. Psychol.*, 1950, 40, 61-73.
17. GAGNE, R. M., & BAKER, K. E. Stimulus pre-differentiation as a factor in transfer of training. *J. exp. Psychol.*, 1950, 40, 439-451.
18. GIBSON, E. J. A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychol. Rev.*, 1940, 47, 196-229.
19. GRANT, D. A., & SCHNEIDER, D. E. Intensity of the conditioned stimulus and strength of conditioning. *J. exp. Psychol.*, 1948, 38, 690-696.
20. HAIMOWITZ, M. L., & HAIMOWITZ, N. R. Reducing ethnic hostility through psychotherapy. *J. soc. Psychol.*, 1950, 31, 231-241.
21. HOVLAND, C. I., LUMSDAINE, A. A., & SHEFFIELD, F. D. *Experiments on mass communication*. Princeton: Princeton Univ. Press., 1949.
22. JOHNS, E. H., & SUMNER, F. C. Relation of the brightness differences of colors to their apparent distances. *J. Psychol.*, 1948, 26, 25-29.
23. KRIEDT, P. H., STONE, C. H., & PATERSON, D. G. Vocational interests of industrial relations personnel. *J. appl. Psychol.*, 1952, 36, 174-181.



which would be considerably changed by the flux which is science were a second analysis made a few years hence.

If the above paragraph suggests some confusion as well as the healthy, normal, expected changes in a vigorous discipline such as psychology is today, then I think it is accurate in its implication. Indeed, I must spend considerable time pointing out sources of confusion if I am to adequately reflect our science as it exists today. I will make no pretense that this confusion can be ameliorated appreciably; there are too many divergent trajectories of thought to expect this even if it were possible and desirable. The best I can hope to do is identify some of the sources of confusion and disagreement and try to provide some common base by which the disagreements can be described.

*Ideal versus actual explanatory attempts.* Throughout our discussions of problems related to explanatory attempts, it will be wise to keep in mind a distinction between the ideal of explanation and the explanatory attempts undertaken by a "typical" research psychologist. To the philosophers of science, we owe a considerable debt, one portion of which is for their penetrating analyses of the formal aspects of concepts and laws in theoretical systems. By this I mean the interrelationships among concepts and laws, how postulates may mediate law-like predictions through the application of the rules of deductive logic, the role of mathematics in theoretical systems, and so on. These analyses have resulted in a fairly standard conception of what a "good" explanatory system consists. But, to use a trite phrase, it may be a mixed blessing to have this ideal model before us. The model has been largely filtered from the work in the older sciences, notably mathematical physics. This discipline, in terms of empirical and explanatory age, is so much older than psychology that a disservice to psychology may result if we attempt to emulate this model explanatory system before having reached the appropriate "explanatory-readiness age."

Furthermore, it is not an unquestionable assumption that theory-building in psychology must eventually follow the path of the physical sciences as mapped out by the philosophers of science. However, the relative success of the physical sciences with their theories, the commonality of men's minds, and the zeal with which

# *An Overview of Explanation in Psychology*

## INTRODUCTION

In introducing the first chapter, I said that science attempts to bring about a description and understanding of nature. The descriptive phase is commonly identified closely with the research of science. Our discussion of the components of the research situation, operational definitions, and experimental designs were all oriented toward this descriptive or empirical component of science. We turn next to the second aspect of science, understanding. By understanding I mean the explanatory or theoretical efforts of the scientist which so often accompany the research.

While I may speak somewhat glibly of description and explanation as two discrete components, it is obvious that in actual practice no such clear separation is apparent. If science were a static affair, if it consisted only of a classification of stable objects, or even events, such a bifurcation as made above might have more value. This is not the case. There is a constant shifting—a continual interplay between the descriptive and explanatory efforts. Explanation is never ultimate in the mind of a scientist. What may be considered adequate explanation today may be relegated to theoretical purgatory tomorrow. What may seem to some to be a straightforward empirical relation may be raised by a theoretician to the level of a postulate and used (with other postulates) to explain other relationships. Thus, a cross-sectional analysis of the explanatory concepts of a science provides no more than a momentary picture, a picture

which produces such diverse opinions as do problems related to explanation, such as what is theory, when should we have it, and so on. And, I know of no other area which presents the open-minded student with a body of writings so difficult to get one's teeth into and find there something substantial to hang on to. Theories about theories shift about in a most unstable and indecisive manner.

# PROBLEMS IN UNDERSTANDING EXPLANATORY ATTEMPTS

*Confusion in terminology.* To a certain extent I have already avoided the use of the word "theory" simply because its meaning is so ambiguous. The ambiguity can by no means be attributed only to the writings of psychologists, since the philosophers of science and scientists in other disciplines have contributed their share to the chaos. Let me sample a few writings on this topic.

Bergmann, a philosopher of science, but quite close to psychology, says:

...as vague as the customary use of the word "theory" itself (2, p. 337).

Campbell, an English physicist and philosopher of science, writes:

...it will be well to start by explaining in some detail exactly what meaning I propose to attach to the term "theory." I shall not assume at the outset that my use of the word coincides with that generally adopted; indeed, since I shall urge that the general use covers propositions of widely different form and significance, I can expressly disclaim that assumption (7, p. 120).

And, a statement by Stafford, a psychologist:

...but we are still without a precise formulation of what we mean by *theory*. Do we, all or any of us, mean by theory, one, all, or none of the following: implicit definition, speculation, postulate, hypothesis, assumption, correlation or coördination of variables, deductive elaboration, explanation, school or system (22, p. 61)?

Bergmann, this time writing with Spence, the psychologist:

Any attempt then to divide this hierarchy of constructs into sheep and goats, i.e., operational constructs and theoretical constructs, is of necessity arbitrary (3, p. 6).

some theorists in psychology at least try to use physical theories as models for their own theoretical efforts, make it seem likely that sooner or later behavior theory will follow the lead of the older sciences at least as far as it can (cf., Hempel & Oppenheim, 12, for a discussion of this matter; pp. 325 ff.) But, to assess most explanatory attempts in psychology against the formal systems of the philosophers of science (as derived from physical theory through philosophical considerations) would be a waste of our time at present. Such an assessment would show only that most explanatory attempts in psychology just simply do not approach the ideal formal structure demanded of advanced explanatory systems. For example, if one looks at the elegance of an explanatory system as given by the philosophers of science (e.g., Cohen & Nagel, 8), with its formal axioms and deduced theorems, one realizes immediately that it rarely contacts reality as far as explanation in psychology is concerned. A philosopher of science is not a working scientist; he is not faced with the day-to-day problems of the search for some form of explanation in the restricted area in which most scientists work. If it be argued that it is not the job of the *typical* scientist to construct comprehensive theory then we might agree, for very few scientists of any discipline have done so. If it is argued that it is not the job of the typical scientist to worry about explanation in his own limited area of research, again there would be no argument. But, I would simply point out that most scientists *do* engage in some kind of explanatory attempts and it is our task to try to come to some understanding of the thinking of the scientist in making these attempts.

But now, to return to an earlier point, we may ask at what time is a discipline ready for these formalized systems of explanations propounded by the philosophers of science? When does the age of "explanatory readiness" arrive? There are a number of issues relevant to the answers to these questions. I think it will be wise, however, if we delay the discussion of these issues until a little more ground work is laid. It may be said, however, by way of anticipation, that on the basis of the assorted opinions of psychologists who have spoken about this matter, we shall arrive at no satisfactory answers to the questions. Among psychologists I know of no other issue

system of ideas or facts (no matter how small the system) when this system allows for deductions. In a very crude sense deductions follow the pattern: "If this is so, and that is so, then this must be so." It is thus a form of *sylogistic reasoning*, although in mathematical systems it cannot be handled so familiarly. Let me sample two of many writers who suggest that the word "theory" should be used only when the deductive arrangement exists among concepts.

Such organization of empirical laws into deductive systems is the distinguishing characteristic of scientific theories (2, p. 336).

It is almost a platitude to say that every science proceeds, more or less explicitly, by thinking of general hypotheses, of greater or less generality, from which particular consequences are deduced which can be tested by observation and experiment (5, p. ix).

We might then agree that when deductive possibilities exist among terms we have the basis for use of the word "theory." In order to keep the exposition in this and subsequent chapters terminologically immaculate I would be delighted to use such a criterion. But I can't. While what is and what isn't deduction may be crystal clear in the sciences where statements are mathematical in nature, in the theoretical efforts among psychologists where words rather than numbers dominate the science, whether or not deduction has taken place, or indeed, can take place is not easy to determine. Deduction, induction, and what might be called scientific intuition become somewhat confused.

In order that a set of concepts may have deductive power, statements must be made concerning interaction among processes symbolized by the concepts. (An elaboration of the nature and meaning of concepts which may have this deductive power must wait until the next chapter.) These statements are such as those about summation, subtraction, multiplication, etc. of the processes. They seem to occur fairly generally in psychological writing. And, to me, this criterion of interaction comes somewhat close to being a way of deciding whether or not a system has deductive consequences. Nevertheless, I have here again been unable to apply a proposed criterion to my own satisfaction; in some cases it is not at all clear that the supposed or postulated interaction process is necessary for the predictions which are made. In short, I think anyone would

Boring, the psychologist, must have despaired of using the word "theory" in any differentiating sense:

In brief, a generalized description is a theory. This meaning for the word *theory* is admitted by those who discuss the philosophy of science, although for the most part they prefer to limit the term to the more complex cases, the theories that exhibit the interrelationships among abstract concepts. I am insisting on the broader meaning because I am arguing from continuity. I am saying that concepts are created by inductive generalization, that science is made up of confirmed relationships among concepts, not among data, that theory is so pervasive that it penetrates even the observational instant, when the observer decides whether to classify his black-white perception as 10.7 or 10.8 on the ammeter scale (4, p. 175).

When Boring says that he insists on the "broader meaning" he means it, for he lists 15 kinds of psychological and scientific theories which range from simple functional relationships between dependent and independent variables to mathematical models. And he says:

The difference between being theoretical and empirical is mostly a question of how far the process of reification of the construct has progressed (4, p. 172).

But, Boring solves nothing—he may not have intended to—for we do not know where in the process of reification one switches to the use of the word "theory." Apparently this is left to the individual scientist to decide.

I need not multiply the quotations. Let me say only by way of addition that the word "theory" is not the only term in this domain to cause confusion. What may be called axioms by some (8) may be labeled postulates by others (13) and hypotheses by still others (7). What may be termed hypothetical constructs by some writers (16) may be called transcendent hypotheses (14), inferred entities (1) or "fictional concepts" (17). And there are many other sources of conflicts (1).

But now, let me turn back to the word "theory" and ask if there is any degree of agreement which it might be useful to exploit. There may be. By a number of writers the assertion is made (in one form or another) that the term is most appropriately used for a

introducing explanatory attempts. What reasons does the psychologist cite for the way in which he theorizes? Let us examine a number of possibilities as they have been suggested by various writers in psychology.

1. It is a commonly observed fact about the organism that it is not a linear transmission system; its output—measured behavior—is not linearly related to the input of the environment. Seldom do we find straight-line functions for relationships between independent and dependent variables, even though we transform our measuring scales in a number of different ways. This fact suggests (it does not prove beyond doubt) that the transmission system from sense organs through to effectors *does something about the input; it doesn't "attenuate" or "amplify" at a one-to-one ratio.* One function of theory has been said to suggest what the modifying processes are, and more precisely, how they modify. These guesses may be made in physiological terms (and so are intended to suggest actual physiological processes) or they may be stated in strictly psychological terms with no specific physiological mechanisms implied.

This function of theory, while approached somewhat differently from the one indicated above, has been suggested by Spence:

Theoretical constructs are introduced...in the form of guesses as to what variables other than the ones under control of the experimenter are determining the response (20, p. 51).

Just how these guesses may assume acceptable theoretical or explanatory status is a matter for later discussion. For the moment, I wish to pursue Spence's thinking further. He goes ahead to point out that these theoretical constructs are *commonly called intervening variables.* Then, to continue, in Spence's words:

If under environmental condition  $X_1$ , the response measure  $R_1$ , is always the same (within the error or measurement) then we have no need of theory. Knowing that condition  $X_1$  existed we could always predict the response. Likewise if, with systematic variation of the  $X$  variable, we find a simple functional relation holding between  $X$  values and the corresponding  $R$  values we again would have no problem, for we could precisely state the law relating to them. But, unfortunately things are not usually so simple as this, particularly in psychology. On a second occasion of the presentation of  $X_1$ , the subject is very

have some trouble in applying either the criterion of "deductive consequences" or the criterion of "interaction of processes."

While there is a certain amount of futility involved in using words which one cannot be satisfied with, we still must communicate as best we can. Therefore, I shall try to restrict myself in the use of the word "theory" to those situations in which the processes symbolized by the concepts interact so as to permit deductions. Though I have seriously considered dropping the word, I see that doing so would not solve the basic problem. We are continually faced with the term in our readings and we must make the best of it that we can. I must be allowed, therefore, to use the word with some looseness of meaning.

The word "explanation" will be used in a very general sense as indicated below. It will be apparent that theory is one very important way or method of attempting to provide explanation in psychology.

*Purpose of explanation.* Upon one general matter, scientists and philosophers of science have reached nearly universal agreement. It is that the purpose of explanation is to account for the greatest number of facts or observations with the fewest number of principles or assumptions. I have used more alternative words in the above statement to encompass differences in usage among various writers, although I have by no means exhausted the variety of terms that have been used. Nevertheless, I think the objective of explanation is clear. The ultimate in explanation would be two comprehensive principles from which all relationships of the universe logically stemmed. It is perhaps needless to say that such a goal seems remote, indeed. Of far more importance for explanatory efforts in psychology is the fact that we work *within* limited areas *within* psychology and current explanatory attempts must be evaluated as to how well they encompass facts or observations within any one area, no matter how small it may be. In other words, at present the purposes of explanation can be attained within very limited areas of research and be perfectly valid. In the long-run development of our science we hope that these areas will become united by common principles, but progress toward that end may be expected to move at a very slow pace.

Within the general statement of the purpose of explanation a number of more specific reasons have been given for theorizing or



do just what was noted earlier as a generally-agreed-upon purpose of explanation, namely, account for a broad range of facts by as few basic principles as possible.

3. Another reason for introducing some theoretical concepts in psychology has been suggested by Tolman (23), and it is accepted by many theoretical-minded psychologists. Even simple experimental situations involve a great many variables affecting the measured response. This would include all environmental, task, and subject variables. Each class does have many specific variables which influence behavior. Now, it is conceivable that the effect of all these variables could be expressed in a single, gigantic equation. But, Tolman says, it is very difficult, if not impossible, intellectually, to comprehend the implications of such an equation. Rather, it is much easier to conceptualize the subprocesses independently. Thus, he suggests that intervening variables first be thought of as being influenced directly by the independent variables. Several independent variables may influence only one intervening variable (thus producing an initial economy of thought). Then, the several intervening variables may be said to determine another higher-order intervening variable (further producing economy of thought), and this higher-order intervening variable is said to produce the measured behavior. Those who have worked with such formulations do indeed attest that conceptualization is more easily attained by initially breaking down the total mass of interrelationships into simpler equations. Hull (13) also indicates that the use of subequations for the first-order intervening variable (one related directly to stimulus variable) makes the problem of deriving a quantitative index for that variable easier. This index would reflect the influence of the several manipulable variables which are said to "operate on" the intervening variable.

4. We have said that the major purpose of explanation is to account for the greatest number of facts with the fewest number of assumptions. If a theoretical or explanatory principle will account for a number of what might at first appear to be diverse facts, a certain heuristic value occurs as a by-product. It has often been said that knowledge is more than a mere collection of facts. I do not care to discuss the meaning of knowledge, but I do wish to point out that as a science develops at the empirical level the number of established

likely to exhibit a different magnitude of response, or in the second example there may be no simple curve discernible between the two sets of experimental values. It is at this point that hypothetical constructs are introduced and the response variable is said to be determined in part by  $X_1$ , and in part by some additional factor, or factors... (20, p. 51).

In my opinion, Spence's reason for theory as given here makes a very weak case. His first illustration indicates simply that there is an unreliable phenomenon, for he says that  $X_1$  does not result in a consistent response. If  $X_1$  does not produce a consistent response, i.e., if the reliability of the phenomenon is not established, then there is nothing to theorize about. (I am sure Spence does not mean in this case that theory is introduced to account for unreliability.) If he means that on successive presentations of  $X_1$  the response changes in a simple regular fashion (as it might in a learning situation) then these regular changes become as predictable as "no change" and on his own premise no theory needs to be introduced. And the complexity of the relationship, if it is reliable, does not change the ability to predict. In short, I find it very difficult to accept this opinion as to why theory is introduced. But, perhaps Spence's opinion has changed, for in a later publication we find a somewhat different reason being advanced for theorizing.

2. Referring to the field of learning, Spence notes that if we are dealing with one response measure in a single type of experimental situation there might possibly be no need for theory because it is feasible that a single mathematical equation would fit the various curves of learning found in the situation when different variables are manipulated. But, if a number of response measures are used in several different experimental learning situations, several different types of learning curves may be found relating the response measure to the independent variables.

Confronted with such a state of affairs, the theory-oriented psychologist has attempted to integrate these isolated, particular sets of laws into a more comprehensive system of knowledge by means of his theoretical formulations (21, p. 153).

It is clear that in this statement of one purpose of theory Spence is suggesting that within the area of learning, theoretical attempts may

apparent without the theory. I do not think we should take this statement (and similar ones) as a methodological axiom. And, the fact that the theory suggests research does not mean that it is automatically significant research (significant in the sense that the discipline is advanced more rapidly than would have been the case if the research had not been predicated on the theory). I have written elsewhere about this matter (24). It has seemed to me that a great deal of effort has been wasted in attempts to test theoretical disagreements in some areas of learning, as typified by the latent-learning studies. The research has tended to be bitsy-type research in which the end product hasn't settled theoretical matters and often has not left us with the sound sorts of relationships between variables which supersede any theory. Skinner has recently commented on this matter of research based on theory as follows:

Research designed with respect to theory is also likely to be wasteful. That a theory generates research does not prove its value unless the research is valuable. Much useless experimentation results from theories, and much energy and skill are absorbed by them (19, p. 194).

Brogden comments in a similar vein.

A theory may organize the results of many researchers, it may bring new relations to light, or it may serve as a catalyst for fruitful experimentation. On the other hand, a theory . . . may impede advancement seriously. It may fail to consider existing experimental evidence that does not support it; it may encourage research to proceed in non-productive channels; or it may define problems verbally that cannot be attacked experimentally (6, p. 224).

It would seem then, that in evaluating the usefulness of theory, we would not only need consider whether or not the theory instigates research which would not have been done, but also whether the research so prompted can be advantageously assimilated into the body of knowledge. In psychology, at least, we have no grounds for accepting theory uncritically as a magical stimulator of research; neither do we have a right to condemn it simply because in the opinion of some it has been wasteful in certain restricted areas. Clearly, there are differences of opinion on the issue, and that is the point to be kept in mind at the present time.

facts can easily become so great that it is beyond the capacity of the human mind to assimilate them under the conditions of our culture. But, if a theoretical principle or principles can account for these facts logically, and if we remember the rules of logic, we can often "cue off" the detailed fact by working out the implications of the theoretical principles. It may also be noted that it is this very characteristic of a "good" theory which allows for the prediction of new facts, a point to be discussed later. But, even this "memory function" of a theory has certain dangers, for relevant facts not predicted by the theory, or contrary to it, may have a tendency to slip away. One may wonder if this is what may have happened to theorists who have been accused by their critics of "overlooking" evidence contrary to their theory. I do not wish to justify the overlooking of such data, but I would be interested in understanding the mechanisms which produce it.

5. Another purpose of theorizing has been stated as "good" theory generates research. Actually, this statement may have two meanings. In the first place one who develops a theory or gets attached to someone else's theory may get thoroughly ego-involved in it. This may strongly motivate to do research to "prove" the theory. If the research is sound, and if the relationships discovered stand by themselves without regard to the theory which instigated them, then I think we would agree that the personal relationship between the scientist and his theory may have had a beneficial effect for science as a whole. I do not know how many psychologists would not do research if they were not motivated by a strong affinity for some theory; it may be many or it may be few. But, it would be misleading to suggest that ego-involvement in a theory is entirely beneficial just because it gets research done. This same ego-involvement may introduce blind spots for the significance of data not directly relevant to the theory. Important facts, acquired as by-products of theory testing, may be ignored. And, in view of what we know about the effect of motivation on behavior, it is possible that theory involvement could lead to a distortion in perceiving and reporting data. Therefore, we must not take for granted that theoretically motivated research behavior can do naught but good.

The second meaning to the statement that theory spawns research is that the theory itself suggests research which would not have been

ful to talk about theoretical readiness for psychology as a whole, for certainly it appears that our formalized theory will develop within areas.

I have intimated that theoretical readiness depends upon the acquisition of a stable body of phenomena and relationships between and among independent variables and these phenomena. In short, the theoretician must have something to theorize about; the theory must of necessity cope with a fundamental body of knowledge if it is to be useful at the integrative level. Such a body of knowledge transcends any particular theory and yet seems necessary before serious formalized theoretical undertakings. This has been true in the sciences which now have highly developed theories. Whittaker, the English physicist says:

Let it be frankly admitted that a certain body of knowledge must have been created by the methods of experimental physics before theoretical physics can make a start; the formulae of reflection and refraction must be known before Huygens can devise his Principle to explain them; but when the conceptions of theoretical physics have been introduced, they have a vitality of their own, and an adaptability to fields other than those in connexion with which they were introduced; and it is by them that the unification of isolated experimental results into comprehensive general theories is achieved (25, p. 48).

Now, let us turn to some remarks of psychologists concerning the theoretical readiness in their special areas. The field of learning in psychology has been one which, for some reason, has been most fertile not only for theoretical attempts but for discussion of the role of theories by men working in the area. Concerning theories in learning, Skinner (19) has most recently taken the position that they may not even be necessary. This fact was noted earlier, for Skinner feels that actual retardation can occur in the field of learning by paying too much attention to theory. It may narrow or restrict (although perhaps not reduce) research efforts and not allow for a more free play in systematic exploration of variables. But, even Skinner does not say theory of a kind will never play an important role in the development of the science of learning:

This does not exclude the possibility of theory in another sense. Beyond the collection of uniform relationships lies the need for a

I do not pretend that these are all of the roles which have been attributed to theory in psychology, but they do represent the major ones. It is clear that we have disagreements about some of the specific roles of theory in our science. I suspect we would have near complete agreement on the general statement that the aim of explanation (via theory) is to account for the greatest number of facts with the fewest assumptions. But, the issue still remains as to whether or not our science is in a state of "explanatory readiness" for the more formalized theoretical attempts. Let us turn to a sampling of current opinion on this matter.

### IS PSYCHOLOGY AT AN "EXPLANATORY READINESS" AGE?

It is, of course, ridiculous to try to determine whether or not psychology had become of theoretical-readiness age on, say, June 10, 1955, or that it will become of age on, say, April 30, 1970. As we shall see, theories in psychology could possess the formal structure required of them by the philosophers of science and yet vary greatly in their comprehensiveness, i.e., in the range of behavior phenomena which they encompass. It is not appropriate to talk about theoretical readiness for psychology as a whole because the areas within psychology differ markedly in their stages of empirical development. For example, in the study of sensory processes, particularly audition and vision, the empirical development is probably at a higher level than for any other area of psychology. In the realm of social processes, on the other hand, the accretion of a body of established phenomena and relationships is only in its initial stages.

I think it can be taken for granted that our theories will continue to develop within very limited areas rather than in terms of comprehensive theories of behavior. It may, however, be expected that there will be immigration of theories from one area to another within psychology. That is, if a highly developed theory is attained in one area it may be expected that it will have some usefulness in adjacent areas so that eventually all aspects of behavior may be incorporated within a single system. But, let us not allow our wishful thinking to take us too many generations beyond the present, and perhaps beyond anything we can foresee with any assurance of its attainment. The point I wish to make is that it is not very meaning-

ful to talk about theoretical readiness for psychology as a whole, for certainly it appears that our formalized theory will develop within areas.

I have intimated that theoretical readiness depends upon the acquisition of a stable body of phenomena and relationships between and among independent variables and these phenomena. In short, the theoretician must have something to theorize about; the theory must of necessity cope with a fundamental body of knowledge if it is to be useful at the integrative level. Such a body of knowledge transcends any particular theory and yet seems necessary before serious formalized theoretical undertakings. This has been true in the sciences which now have highly developed theories. Whittaker, the English physicist says:

Let it be frankly admitted that a certain body of knowledge must have been created by the methods of experimental physics before theoretical physics can make a start; the formulae of reflection and refraction must be known before Huygens can devise his Principle to explain them; but when the conceptions of theoretical physics have been introduced, they have a vitality of their own, and an adaptability to fields other than those in connexion with which they were introduced; and it is by them that the unification of isolated experimental results into comprehensive general theories is achieved (25, p. 48).

Now, let us turn to some remarks of psychologists concerning the theoretical readiness in their special areas. The field of learning in psychology has been one which, for some reason, has been most fertile not only for theoretical attempts but for discussion of the role of theories by men working in the area. Concerning theories in learning, Skinner (19) has most recently taken the position that they may not even be necessary. This fact was noted earlier, for Skinner feels that actual retardation can occur in the field of learning by paying too much attention to theory. It may narrow or restrict (although perhaps not reduce) research efforts and not allow for a more free play in systematic exploration of variables. But, even Skinner does not say theory of a kind will never play an important role in the development of the science of learning:

This does not exclude the possibility of theory in another sense. Beyond the collection of uniform relationships lies the need for a

ful to talk about theoretical readiness for psychology as a whole, for certainly it appears that our formalized theory will develop within areas.

I have intimated that theoretical readiness depends upon the acquisition of a stable body of phenomena and relationships between and among independent variables and these phenomena. In short, the theoretician must have something to theorize about; the theory must of necessity cope with a fundamental body of knowledge if it is to be useful at the integrative level. Such a body of knowledge transcends any particular theory and yet seems necessary before serious formalized theoretical undertakings. This has been true in the sciences which now have highly developed theories. Whittaker, the English physicist says:

Let it be frankly admitted that a certain body of knowledge must have been created by the methods of experimental physics before theoretical physics can make a start; the formulae of reflection and refraction must be known before Huygens can devise his Principle to explain them; but when the conceptions of theoretical physics have been introduced, they have a vitality of their own, and an adaptability to fields other than those in connexion with which they were introduced; and it is by them that the unification of isolated experimental results into comprehensive general theories is achieved (25, p. 48).

Now, let us turn to some remarks of psychologists concerning the theoretical readiness in their special areas. The field of learning in psychology has been one which, for some reason, has been most fertile not only for theoretical attempts but for discussion of the role of theories by men working in the area. Concerning theories in learning, Skinner (19) has most recently taken the position that they may not even be necessary. This fact was noted earlier, for Skinner feels that actual retardation can occur in the field of learning by paying too much attention to theory. It may narrow or restrict (although perhaps not reduce) research efforts and not allow for a more free play in systematic exploration of variables. But, even Skinner does not say theory of a kind will never play an important role in the development of the science of learning:

This does not exclude the possibility of theory in another sense. Beyond the collection of uniform relationships lies the need for a



I do not pretend that these are all of the roles which have been attributed to theory in psychology, but they do represent the major ones. It is clear that we have disagreements about some of the specific roles of theory in our science. I suspect we would have near complete agreement on the general statement that the aim of explanation (via theory) is to account for the greatest number of facts with the fewest assumptions. But, the issue still remains as to whether or not our science is in a state of "explanatory readiness" for the more formalized theoretical attempts. Let us turn to a sampling of current opinion on this matter.

#### IS PSYCHOLOGY AT AN "EXPLANATORY READINESS" AGE?

It is, of course, ridiculous to try to determine whether or not psychology had become of theoretical-readiness age on, say, June 10, 1955, or that it will become of age on, say, April 30, 1970. As we shall see, theories in psychology could possess the formal structure required of them by the philosophers of science and yet vary greatly in their comprehensiveness, i.e., in the range of behavior phenomena which they encompass. It is not appropriate to talk about theoretical readiness for psychology as a whole because the areas within psychology differ markedly in their stages of empirical development. For example, in the study of sensory processes, particularly audition and vision, the empirical development is probably at a higher level than for any other area of psychology. In the realm of social processes, on the other hand, the accretion of a body of established phenomena and relationships is only in its initial stages.

I think it can be taken for granted that our theories will continue to develop within very limited areas rather than in terms of comprehensive theories of behavior. It may, however, be expected that there will be immigration of theories from one area to another within psychology. That is, if a highly developed theory is attained in one area it may be expected that it will have some usefulness in adjacent areas so that eventually all aspects of behavior may be incorporated within a single system. But, let us not allow our wishful thinking to take us too many generations beyond the present, and perhaps beyond anything we can foresee with any assurance of its attainment. The point I wish to make is that it is not very meaning-

questions we can be open-minded and are likely to let the answer to one question influence the nature of the next. It is this way that we get acquainted with our universe. However, when we predict we show our maturity, and we can even determine a scientist's success by calculating his percentage of correct predictions. But we must not act more mature than we really are (18, p. 54).

Thus Maier's position is essentially "go easy." Seven psychologists who analyzed five major theoretical attempts in the field of learning have this to say:

On the one hand, we appreciate the need for a common theoretical structure to facilitate the ordering and application of our knowledge of learning. On the other, we recognize the complexity of the material which must be handled by a behavior theory. Vigorous individual attempts at theory construction along a wide variety of fronts are probably not only desirable but necessary for continued progress in this area (10, p. xiii).

MacKinnon, whose work has been largely in the area of personality, says:

...I shall now express my opinion that in personality research our theorizing and building of models have outrun activities more intimately concerned with observation and data collecting. Our greatest need for the more adequate study of personality is systematic observation and systematization of the data we collect, and this, I submit, is something more than theorizing (17, p. 141).

Koch, writing in 1950, states his belief as follows:

We must start with the humiliating acknowledgment that psychology is in a *pre-theoretical stage*, and that the central problem of the fundamental psychologist is not what doctrine to embrace or concoct, but simply to assay, realistically, how psychology can be made to move towards adequate theory (15, p. 298).

Finally, let me cite George, an English psychologist. In proposing a reduction of linguistics of psychology to logical forms, and in speaking of theory in general, he somewhat petulantly condemns the strict empiricist as follows:

This approach, it is hoped, will especially be brought to the notice of those narrow experimentalists who repeatedly call for experiment,

formal representation of the data reduced to a minimal number of terms. A theoretical construction may yield greater generality than any assemblage of facts. But such construction will not refer to another dimensional system and will not, therefore, fall within our present definition (19, p. 215-216).

We do not seem to be ready for theory in this sense (19, p. 216).

Spence (21) makes it clear that the first objective of the scientist is to establish a set of laws relating the independent and dependent variables. Only when some precision is obtained for such relationships is the scientist ready to develop a formal theoretical system, albeit a system that explains a very narrow range of behavior. The fact that Spence has done considerable theorizing of a type in certain areas of learning suggests that he must believe that the state of empirical knowledge warrants such theorizing. One cannot be sure of this, however, for again we must keep distinct the formal theoretical systems and other explanatory attempts; Spence's theorizing does not easily fit the formal systems idea. Hull (e.g., 13) has been the most ardent advocate of the formal theoretical systems in the field of learning and more than any other psychologist has accompanied his ardor with the construction of the systems. How "good" or "useful" these systems are is another matter.

Maier takes an in-between position with regard to the readiness age for theorizing in psychology of learning in an article called: "Premature Crystallization of Learning Theory."

I personally feel that an interest in theories is desirable for the development of science because theories help us organize facts and they help us to ask good research questions. However, an interest in theories can become a liability if it prevents us from exploring certain kinds of relationships or causes us to ignore facts if they do not fit the theory with which we identify ourselves. When these things occur, the theory becomes an attitude and ideas become good or bad rather than right or wrong.

Perhaps we are somewhat overambitious and have assumed that psychology is more advanced than facts warrant. We seem to want a learning theory that works not only for all learning situations but also for all behavior. We seem to want to predict, to do research by stating hypotheses, and seem no longer to be content with asking questions of the universe and getting our answers through research. When we ask

present; others have both. Some can carry their theories lightly (e.g., 9), others never shed them. For those who have little interest in theory construction but do have interest in research, it is evident that the present state of theory in psychology, and thinking about theory, puts little restriction on their efforts. That is, there is a wide range of problems in psychology which need systematic exploration of the sort that derive direct relationships between dependent and independent variables. Such lawful relationships stand as a basic contribution to our science. If one has no interest in theory he need only fit his research into the empirical framework of the area as it stands at the moment. Others who have interest in theory will sooner or later fit his findings into a theoretical system. Let us recognize the fact of individual differences in abilities and interest and not expect every scientist to be able to do all the things which science in its totality is. Important discoveries have been made and will continue to be made by asking simple questions about the functioning of nature. Likewise, important contributions have been made and will continue to be made by theoreticians as they organize apparently diverse facts. The history of science gives no basis for the disparagement of either the systematic empiricist or the theoretician at any stage in the development of a science.

#### PLAN FOR THE FOLLOWING CHAPTERS

Having attempted to suggest some of the problems we face, some of the disagreements that dominate the over-all picture of explanatory attempts in psychology, I shall now indicate the approach which will be taken in the following chapters. I said earlier that it is rather futile to analyze explanatory attempts in psychology by comparing them with what I have called the ideal formalized structure recommended by philosophers of science as a result of their analyses of highly developed physical theories. We are interested primarily in how theories get started and how they grow. Looking at the formalized system, the *fait accompli*, might help us in setting our sights for the distant future but it does not at all reflect the agonizing work of the many many scientists which went into the making of the system; it does not represent the false starts and the inelegance of theorizing which crop up almost everywhere in the efforts of the

and decay theory, and continue to stumble through the maze of science erroneously believing themselves to be dealing wholly with facts, and never with linguistics (11, p. 232).

Now, let me quickly add, that the issue of whether we should or should not have theory, or the issue of whether if we are to have it, just how soon we are going to have it, is not going to be settled by the kind of opinion poll, of which I have given an inadequate sample above. But even this poll reflects the diversity of opinion and that is its purpose. I think it is well to remind ourselves continually that at our present stage of development in psychology, there *are* very diverse opinions on these matters. I hold the opinion, developed from talking with psychologists at a number of universities, that many Ph.D. candidates have been led to believe that unless a piece of research is predicated on at least one hypothesis, developed somewhat formally from some assumptions, it can have little value. Such beliefs are provincial in the sense that they do not reflect the thinking of a number of respected scientists in psychology today who feel that we should not allow ourselves to get immersed in theory at the present stage.

But still, to repeat, no opinion poll will settle this issue; the best that could be hoped for is that from it some tolerance would result. Whether or not we are ready for formalized theory in certain areas of psychology can only be judged by the success such theories may have in working toward an organization of diverse facts under a few basic principles. The experts in each area will have to be the judges of this matter. For some of the more general theories in the field of learning the current evaluation is not too encouraging, although the same writers who have heavily criticized the theoretical attempts call for more—but better—attempts (10).

#### THE INDIVIDUAL SCIENTIST AND THEORY

One of the basic observations of human behavior is that individuals differ. These individual differences exist among psychologists. Some psychologists develop no interest in theory; others do. Some do not have the motivation or ability to cope with theoretical matters; others do. Some have not had the training even if the motivation is

ignored or the meaning and intent of the concept will be too greatly masked.

# REFERENCES

1. BECK, L. W. Constructions and inferred entities. In H. FEIGL & M. BROOBECK (Eds.) *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953.
2. BERGMAN, G. Outline of an empiricist philosophy of physics (concluded). *Amer. J. Physics*, 1943, 11, 335-342.
3. BERGMAN, G., & SPENCE, K. W. Operationism and theory construction. *Psychol. Rev.*, 1941, 48, 1-14.
4. BORING, E. G. The role of theory in experimental psychology. *Amer. J. Psychol.*, 1953, 66, 169-184.
5. BRAITHWAITE, R. B. *Scientific explanation*. London: Cambridge Univer. Press, 1953.
6. BROGDEN, W. J. Some theoretical considerations of learning. *Psychol. Rev.*, 1951, 58, 214-229.
7. CAMPBELL, N. R. *Physics: The elements*. Cambridge Univer. Press, 1920.
8. COHEN, M. R., & NAGEL, E. *An introduction to logic and scientific method*. New York: Harcourt, Brace, 1934.
9. DALLENBACH, K. M. The place of theory in science. *Psychol. Rev.*, 1953, 60, 33-39.
10. ESTES, W. K., KOCH, S., MACCORQUODALE, K., MEEHL, P., MUELLER, C., SCHOENFELD, W., & VERPLANCK, W. *Modern learning theory*. New York: Appleton-Century-Crofts, 1954.
11. GEORGE, F. H. Formalization of language systems for behavior theory. *Psychol. Rev.*, 1953, 60, 232-240.
12. HEMPEL, C. G., & OPPENHEIM, P. The logic of explanation. In H. FEIGL & M. BROOBECK (Eds.) *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953.
13. HULL, C. L. *Principles of behavior*. New York: D. Appleton-Century, 1943.
14. KNEALE, W. Induction, explanation, and transcendent hypothesis. In H. Feigl & M. Brodbeck (Eds.) *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953.
15. KOCH, S. Theoretical psychology, 1950: An overview. *Psychol. Rev.*, 1951, 58, 295-301.
16. MACCORQUODALE, K., & MEEHL, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychol. Rev.*, 1948, 55, 95-107.

scientist. The bulk of the explanatory attempts in psychology today, the explanatory attempts of the typical psychologist interested in the matter, deals with only two or three explanatory concepts and a limited area of behavior. It thus seems to me that to reflect most faithfully the explanatory attempts in psychology, we must deal extensively with single concepts and relationships among a few concepts. And our prime purpose is to comprehend, not exhort.

So our initial and major effort in the material to follow will be directed toward single concepts. That is, taking a single concept, we must discuss its characteristics, what it is intended to do, how it differs from other concepts, opinions concerning how it should be used, and how it is related to other concepts which may be said to constitute an explanatory system. The task is not a simple one, for (a) a theory-like concept may be used in different ways by different writers; (b) the way in which a particular concept is being used by a writer is not always clear, and (c) the ways in which concepts differ are numerous. This presents, then, a somewhat unpalatable task, for surely injustice will be done some writers and some concepts. And, although I shall deal only with *some* of the differences among concepts, and shall try to make the differences clear enough by example and pointedness, it goes without saying that sharp lines of distinction will be an exception.

The approach will be quite simple. I try to take the viewpoint of the individual research worker in the laboratory and look at the problems of explanation which confront him as he tries to "make sense" out of the data in his own limited area of research. We then try to analyze what he comes up with and occasionally speculate on his thoughts as he tries to move from data toward explanation.

The following chapter deals exclusively with analyses of individual concepts. In this chapter I must again beg indulgence for some repetition of materials occurring in Chapter 3 (Operational Definitions). It was not my intention there to fully set forth the differences which actually prevail in the usage of operationally defined concepts. I shall try to correct these omissions in the following chapter. When dealing with the status and characteristics of individual concepts I shall try to avoid as much as possible the interrelationships among concepts although in some cases this must not be

## *Some Characteristics of Concepts*

Because concepts with widely different characteristics eventually enter into explanatory attempts, I must emphasize these differences in concepts as concepts before I consider how they may be used in explanations. This objective is the sole intent of this chapter.

I do not feel that there is a good, single descriptive dimension along which I can order the concept analysis which is to be made. Differences among concepts are multidimensional. In a very loose sense I shall proceed along a dimension of abstraction, by which I mean one based upon how far the concept is removed from immediate data. There should be some way, as a sheer matter of convenience, for designating modal points along this rough dimension. Being blessed with extraordinary inventiveness I have chosen to speak of five points called Level 1, Level 2, Level 3, Level 4, and Level 5.

### LEVEL-1 CONCEPTS

As I indicated near the close of Chapter 3, I did not reflect fully the differences in attitude toward the question about what operational definitions define. This was intentional, in part to keep the discussion tidy and in part because it is my personal bias that operational definitions should be largely concerned with defining behavioral phenomena. In the present discussion I will try to correct this one-sidedness arising from my intolerance.

In Chapter 3, I insisted that operational definitions are concerned with behavioral phenomena; we define these phenomena by specifying the measuring operations required. However, certain writers say that operational definitions are given to independent variables. In my discussion, the definition of independent variables was a part of the operational definition of a phenomenon. I limited operational defini-



17. MACKINNON, D. W. Fact and fancy in personality research. *Amer. Psychol.*, 1953, 8, 138-146.
18. MAIER, N. R. F. The premature crystalization of learning theory. In *The Kentucky symposium*. New York: Wiley, 1954. Quotations reprinted with permission of John Wiley & Sons, Inc.
19. SKINNER, B. F. Are theories of learning necessary? *Psychol. Rev.*, 1950, 57, 193-216.
20. SPENCE, K. W. The nature of theory construction in contemporary psychology. *Psychol. Rev.*, 1944, 51, 47-68.
21. SPENCE, K. W. Mathematical formulations of learning phenomena. *Psychol. Rev.*, 1952, 59, 152-160.
22. STAFFORD, J. W. Fact, law, and theory in psychology. *J. gen. Psychol.*, 1954, 51, 61-68.
23. TOLMAN, E. C. The intervening variable. In M. MARX (Ed.) *Psychological theory*. New York: Macmillan, 1951.
24. UNDERWOOD, B. J. Learning. *Ann. Rev. Psychol.*, 1953, 4, 31-58.
25. WHITTAKER, E. Eddington's principle in the philosophy of science. *Amer. Scient.*, 1952, 40, 45-60.

*second*) are operationally defined by the physicist. And, if one wishes to trace the history of such definitions it can be seen that all are based on the *discriminatory response* of the human observer—the scientist (17). But, let us pursue these matters no further. There are two points which, by way of summary, I wish to make. First, it is necessary to specify what one means by his independent variable, and such specifications are sometimes called *operational definitions*. Second, certain words which may be used to summarize such specifications by some are used in quite different ways by others. The Level-1 concept as used here refers to the specification of the independent variable without immediate reference to the behavior of the subject.

### LEVEL-2 CONCEPTS

We have seen that explanatory attempts are undertaken when there is some body of reliable knowledge. For psychology I have spoken of this body of knowledge simply as reliable phenomena, including thereunder reliable relationships between dependent and independent variables of all degrees of precision as well as relationships among dependent variables. These constitute the data which we try to explain. Usually such data are gotten by research, although this is not necessary as long as we can be sure of the reliability of the phenomena. Thus, it wouldn't require a Latin-square design to determine that apples consistently fall to the ground rather than into the sky. In psychology most behavior in which we are interested is somewhat more subtle than a falling apple and we are commonly faced with the explanation of research-derived relationships. The notion for an explanation may come from casual observation and it may be supported ostensibly by drawing attention to incidental observations, but, for the most part critical explanatory attempts start and end with data derived from research.

These considerations indicate that to fully understand the use of concepts in psychology we must return to our discussion of the characteristics of concepts employed to summarize operations used in defining phenomena. We might think that we could make short work of this and immediately proceed to concepts which have explanatory-like status. This is not so, for again, I must apologize

tions to an if-if-then type of statement. Nevertheless, because certain writers speak of independent variables as being defined operationally, I am recognizing their position here and am calling these variables Level-1 concepts. It is especially important that this be done at this point for sometimes words which are used to summarize operations by some writers are used in quite another way by others. Let us look at a few illustrations.

The term *extinction* is sometimes used in conditioning to indicate simply the removal of the unconditioned stimulus. That is, it refers only to an activity of the experimenter, not the activity of the experimenter *and* the resulting behavior of the subject. It thus specifies a change in the experimental conditions and that is all it specifies. The word *reinforcement* is sometimes used to indicate that the experimenter gave the animal food after a correct response; by others it is used to include both the giving of the food and the resulting change in behavior.

The experimenter may say that he is operationally defining *deprivation time* (an independent variable) by number of hours since feeding. *Cycles per second* is defined as number of undulations of the sign wave in one second. *Cortical lesion* may be specified minutely in terms of surgical-techniques used, exact site of lesion, and so on.

Level-1 concepts, therefore, refer to activities of the experimenter in specifying what he means by a particular term used as a name for an independent variable; such definitions do not include behavioral aspects of the subject; they do not define behavioral phenomena in the sense that I have used the term in discussing operational definitions. Of course, when such terms are used to indicate what is meant by the independent variable, their meaning must be made perfectly clear. I simply have not included such specifications as operational definitions independent of the behavior of the subject. This is quite an arbitrary decision on my part. I may note, however, that many independent variables become such only after being operationally defined behaviorally. For example, the class of operations which I called *Scaling Operations* in Chapter 3 may result in a reliable dimension which is subsequently manipulated in an experiment as a part of the definition of another behavioral phenomena. The physical scales which are used in psychological investigations (e.g., *cycles per*

cepts do not relate to any theory-like implications since not one whit of explanatory prejudice resides in the definition.

I should note again, as in the chapter on operational definitions, that all operationally established phenomena in psychology do not have a name. For example, the fact is that differences in performance that result from massed and distributed practice do not have a name, but as a phenomenon would usually be called a Level-2 concept. In deference to brevity, my illustrations will make use of phenomena to which names are usually assigned.

2. *Figural after-effect.* I suspect any phenomenon named with the terminal word "effect" will be a Level-2 concept. The implication is clearly that of being a name for a behavioral phenomenon without any extra-operational cause implied. In the case of figural after-effect the operations specify that if a particular figure is fixated for a short period of time, and if then a test figure (with certain specifications) is observed, certain distortions will be reported, the distortions being delineated by comparison with a control in which the original figure was not fixated. The critical feature of the operations is the fixation of the original figure and again, this may be thought of as the cause for the subsequently measured distortion. But, nothing in the definition implies or even intimates a cause over and above this fixation.

3. *Experimental extinction.* In illustrating Level-1 concepts I said that the present term is used in a Level-1 sense sometimes. However, I think that the majority of writers think of this as a behavioral phenomenon of the subject. Furthermore, it is defined as a function of the critical variable (removal of the unconditioned stimulus) and no other causal mechanism is implied; the change in behavior resulting from these particular operations is experimental extinction. Again, it is quite another matter to suggest some process or mechanism resulting from the removal of the unconditioned stimulus which is said to account for or explain extinction.

I need not extend the illustrations; it is enough for us to recognize that Level-2 concepts are as empirical as it is possible to make a behavioral phenomenon via operational definitions. These concepts represent phenomena which have a strong character of "thingness" or "point-at-ableness." While we of course recognize that a phenomenon does not exist apart from the organism, an operational definition at Level 2 might seem to suggest that the phenomenon identified

for not fully representing in Chapter 3, the differences in attitude which exist toward operationally defined concepts. As a matter of fact it is this difference in attitude toward operationally defined concepts which, in my opinion, has created a most perplexing confusion (if confusion can be anything but perplexing). This confusion will come out most clearly when I subsequently contrast Level-2 concepts with Level-3 concepts. So first, I must indicate what I mean by Level-2 concepts.

A Level-2 concept is one which summarizes the operations used to define a phenomenon and therefore merely identifies the phenomenon. I call it *phenomenon identification* or *phenomenon naming*. The definition of the phenomenon implies not one thing about a causal process or condition over and above the operations *per se*. My presentation of operational definitions in Chapter 3 was entirely of this nature. That this presentation was by no means reflective of current practice in psychology will become evident (and I repeat myself) in discussing Level-3 concepts. For now, however, let me give you some illustrations of Level-2 concepts to jog your memory of Chapter 3.

1. *Reminiscence in motor learning*. To define adequately this phenomenon we use the E/C, S-R type of operational definition. To state the definition in general terms we would say that if two groups are given trials on a motor task, and if after a specified number of trials the first group is given a short rest and the second not, and if the performance of the first group is superior to the second after rest, reminiscence is defined. Difference in the performance of the two groups after the rest of the first group is reminiscence. The phenomenon has, in such definitions, almost a "point-at-able" status. Nothing is implied in the definition about any causal factor other than the rest pause. The rest pause is the critical independent variable in the defining operations. Now, of course, that an investigator may give such a definition doesn't mean that he doesn't have ideas or notions about processes taking place during the rest pause which may have produced the observed difference. He simply is not letting these ideas enter into his definition. At the strictly empirical level the rest pause is the cause of reminiscence but explanatory attempts may be expected to go beyond such a notation. These Level-2 con-

concepts, but first we need some illustrations of concepts which are commonly given in Level-3 language. Again, for the sake of brevity, I shall not give complete definitions, for the concepts are familiar and we need not worry about the formalities in this particular instance. Some of my illustrations were used in Chapter 3, but there I used Level-2 language since I was at that time trying to avoid the present conflict.

1. *Drive*. I suspect we have no purer case of a Level-3 language than that commonly used to define drive in animals. Drive may be defined by relating differences in deprivation time (such as for food or water) to performance differences (such as activity level). Having done this it would be quite unorthodox to "point at" the performance difference and say "that is drive" (as we would do at Level-2). Rather, we almost universally think of drive as something which *causes* the performance difference; we say that differences in drive *caused* the performance difference. This "something" may be thought of as a purely abstract something or it might be thought of as changes actually taking place in the organism, without specifying in the definition what these are. (The matter of locus or reality of such concepts will be discussed later). Nevertheless, our definition implies something which is changed by changes in deprivation time and this in turn causes the performance difference which we have observed.

2. *Frustration*. In defining frustration our attitude toward what we are dealing with usually puts a process or state (called frustration) "inside" the organism to account for the difference in performance.

3. *Repression*. If this concept were operationally defined successfully, i.e., if reliable performance differences could be obtained as a consequence of unique operations, most psychologists would, I feel confident, think about it as something which caused the performance difference.

These three illustrations indicate phenomena which are defined by E/C, S-R type operations. I shall shortly turn to concepts defined by response identification. It will be remembered that certain phenomena are defined by S-R identification with a physical stimulus scale and some with a psychological scale. The operations are somewhat different from those used in E/C, S-R definitions. I

exists without reference to the organism (the subject). The Level-2 concepts, via their definitions, suggest that we are dealing with cold, hard, empirical findings; the subject is largely ignored. This is in contrast to the Level-3 operational definitions to which I now turn.

### LEVEL-3 CONCEPTS

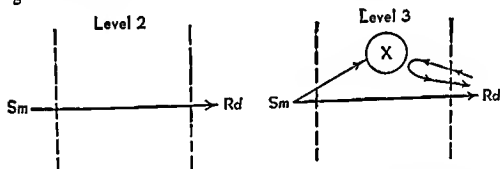
Like Level-2 concepts, Level-3 concepts are operational definitions of phenomena; the only difference comes in the wording of the definition, usually in the final phrase. This difference in wording, I strongly suspect, comes about largely because of differences in the conceptual predilections of the scientist regarding certain phenomena. Level-3 concepts name or identify a phenomenon just as do Level-2 concepts; but, the name is applied to a hypothetical process, state, or capacity as a *cause* for the observations indicating the phenomenon. Thus, if we use the E/C, S-R operational difference, the definition details the operations and then, in one way or another, says: "this difference between the two conditions leads me to infer a process (state, capacity) *causing* the difference, and I shall call this Process X." Level-3 concepts may be thought of as *causal naming* or *causal identification*.

Let me contrast this type of conceptual thinking with that leading to Level-2 concepts. For illustration I will use reminiscence in motor learning. The definition at Level-2 stated that the difference in performance *is* reminiscence. If the definition were recast into Level-3 language, we would say, after indicating the operations: "If a difference obtains between performance of the two groups after rest, I will infer a process which caused the performance difference and I will call this process reminiscence." I do not find that such a definition is customarily given for reminiscence but such *could* be given and when so given it has the basic characteristic I am using to identify Level 3.

The implication of the considerations thus far is that both Level-2 and Level-3 concepts may be, and often are, based on the same formal type of operations; the differences come about because of the way the scientist thinks about what he is dealing with, and this difference in the scientist's thought processes is reflected in the definition. I shall return shortly to some further characterizations of Level-3

locus or position within the organism. Nevertheless, many scientists using Level-3 concepts often find it very difficult *not* to think in terms of a "real" process as opposed to an abstraction existing nowhere except in the scientist's fantasies. (See Kneale, 9, p. 354 and Beck, 2, p. 370 for discussion of this problem in the physical sciences.) It should be clear, therefore, that the vertical bars in the diagram to follow do not represent the soma of an organism unless the reader puts it there. The space between the bars need represent nothing more than the infinite domain of scientific abstraction.

In defining concepts operationally by S-R type definitions we have specified stimulus manipulations (*Sm*) and specified response differences (*Rd*). Level-2 concepts avoid the idea of a state or process and so the concept is defined by referring directly to the relationship between *Rd* and *Sm*. Level-3 concepts name a process or state as causing *Rd* and this process or state is related directly to *Sm*. Diagrammatically, these may be depicted as in the accompanying figure.



For the Level-3 concept I have drawn bidirectional arrows between *X* and *Rd*. For, in using these concepts the investigator infers a state or process *only if* a reliable difference in response occurs and then says that this difference is caused by the state or process (*X*). Differences in *X* are in turn caused by *Sm*. If this sounds to you like scientific doubletalk, then at this point I must agree. And, it should be mentioned that Level-3 definitions do not always make circularity of the inference so obvious as I have made it here, but it is inevitably present. For example, suppose the definition ends with these words: "...if a reliable difference is obtained I shall infer a process which will be called *X*." This definition leaves off the tag



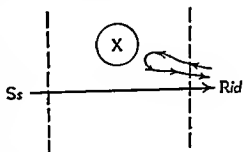
see no reason why these other two types of S-R identification procedures could not be used for defining Level-3 concepts. But, usually this is not the case. Rather, these concepts are handled by what I have called the Level-2 definitions. Thus, *pitch*, *brightness*, *contrast*, *flicker fusion*, and so on, appearing most commonly in sensory-perception studies, are usually defined at Level 2. Occasionally, having defined the term at Level 2, the writer may slip and talk as if it (the defined phenomenon) is now causing itself.

There are phenomena which are by no means consistently placed at Level 2 or Level 3. For example, it seems to me that the phenomenon of *generalization* is sometimes placed at Level 2 and sometimes at Level 3. Also, *transfer* is ambiguously placed. Sometimes the phrase *transfer effect* is used, which clearly puts it at Level 2. But at other times the word *transfer* implies a state of interaction which causes a given performance difference and thus falls nearer to Level 3. The term *closure*, is sometimes used to indicate a process or cause of a phenomenon and sometimes to indicate the phenomenon itself. However, such ambiguities are not my primary concern at this point. The major point I wish to make is that there are concepts, based on the same formal type of operations, which are "thought about" differently by psychologists. The distinction between Level 2 and Level 3 is intended to reflect this difference in scientists' thought processes.

#### FURTHER DIFFERENCES AND SIMILARITIES AMONG LEVEL-2 AND LEVEL-3 CONCEPTS

It may be useful to diagram the differences which are involved in these two levels. But, at the same time it should be recognized that there is a certain danger involved in static diagrams. Either at Level 2 or Level 3 the relationships among the various terms in the definitions (that is, the stimulus manipulations, the response difference, and, for Level 3, the hypothetical process or state which produced the difference) need not imply anything about the organism. The hypothetical process in Level-3 concepts need not imply anything except a name for an assumed causal process. This causal process is inferred from the empirical relationship, but it need not have a

this capacity among individuals which caused the difference. To make the diagram identical to that used for S-R Level-3 concepts an arrow would be added between Ss and X. However, in this case the meaning of such a connection would be quite obscure. Would we say that Ss caused the differences in X? No, I think this would be revolting to most psychologists; Ss only enabled us to establish that individuals differ in amount of X. Did the "quantity" of X for a given person result from Ss? I think most would say "no" again since this is really the same



question. A given amount of capacity or state X already "exists;" Ss was merely a vehicle which allowed a demonstration of this. Like the Skinner-box, a pursuit rotor, or a Snellen chart, the paper-and-pencil test (such as an intelligence test) is a means of eliciting behavior. From differences in this behavior elicited by the test, differences in capacities of subjects are defined. Of course, X has its related antecedents (such as heredity in the case of intelligence) but they do not enter into the definition on Level 3.

*Elaboration of Level-2 and Level-3 concepts.* There are two types of elaboration of Level-2 and Level-3 concepts which I wish to discuss. One may be thought of as *operational identification*, the other, *stimulus variable elaboration*. While these are not completely independent, it will simplify our discussion to treat them separately.

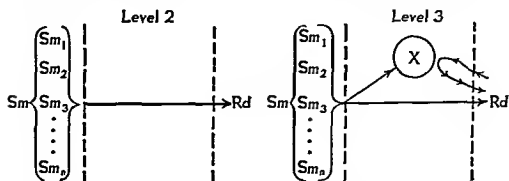
In Chapter 3, I noted that several specific operations may be classed together if they meet the criterion of the class. Thus, *frustration* is defined as resulting from blocking of goal-directed behavior; *retroactive inhibition* as inserting an interpolated task between original learning and recall. *Drive* in general could be defined as resulting from deprivation of biological needs (food, water, sleep, and so on). All that a general definition asserts is that there must be a reliable performance difference (E/C) for the phenomenon to be defined. Now, it is quite apparent that differences in the specific operations may produce differences in the "amount" of the phenomenon observed. *Operational identification* refers simply to the

end of the definition which would say: "X caused the difference." But, this causal inference is implicit for if no reliable response difference were found X would not have been inferred; X must be, in a manner of speaking, responsible for the performance difference. In short, it would appear that Level-3 concepts have a certain amount of risk about them which is not present on Level 2. But, let us delay further evaluation until we have had an opportunity to see how such concepts may be elaborated. I wish to turn now to the status of simple response-defined concepts.

*Simple response-defined concepts.* In Chapter 3, it was pointed out that this class of operationally-defined concepts is well typified by concepts resulting from the use of paper-and-pencil tests. Take the case of *intelligence*. If intelligence is defined at Level 2 we indicate that there must be reliable individual differences on a specified test. A given individual's intelligence is defined as the score made on the test specified (related, of course, in some fashion to scores made by other individuals). Very few psychologists use such definitions; rather, the response measure (score on test) is used to infer a hypothetical state or capacity which is called intelligence. Intelligence is responsible for the score on the test. And, I think that most psychologists are almost compelled to think of this as some capacity which "really" exists in the organism, although I repeat that such thinking is not demanded by the operations. So, then, I am asserting that intelligence is most commonly developed as a Level-3 concept. I would further assert that many other subject capacities are conceptualized in this fashion, e.g., *mechanical aptitude*, *introversion*, *anxiety*.

Let me first remind you that concepts such as intelligence are defined through the use of a static stimulus situation. That is, all subjects are treated in the same fashion—there is no active stimulus manipulation as in the case of S-R defined concepts. To define the concept all that is needed are reliable individual differences in response to this static stimulus situation. I will call the static stimuli  $S_s$ , and the individual differences  $R_{id}$ . In most simple form, the depiction would be as shown in the accompanying diagram. What this diagram means is that if  $S_s$  produce reliable individual differences, an inference is made that there exists differences in the "amount" of

for the specific operations fitting under the general operations, and we can picture the two levels as in the accompanying diagram.



Operational identification also takes place in response-identified concepts. Actually, this identification is what I have called complex response identification in Chapter 3, and is typified by factor analytic procedures. Factor-analytic attempts may be, and usually are, preceded by logical-rational considerations or the formulation of hypotheses about the composition of subject capacities. I think it will be well to quote Thurstone on this matter:

In the light of a good deal of experience with the factorial methods, we should be able to give students a few practical suggestions. In the Psychometric Laboratory at Chicago, we spend more time in designing the experimental tests for a factor study than on all of the computational work, including the correlations, the factoring, and the analysis of the structure. If we have several hypotheses about postulated factors, we design and invent new tests which may be crucially differentiating between the several hypotheses. This is entirely a psychological job with no computing. It calls for as much psychological insight as we can gather among students and instructors. Frequently we find that we have guessed wrong, but occasionally the results are strikingly encouraging. I mention this aspect of factorial work in the hope of counteracting the rather general impression that factor analysis is all concerned with algebra and statistics. These should be our servants in the investigation of psychological ideas. If we have no psychological ideas, we are not likely to discover anything interesting, because even if the factorial results are clear and clean, the interpretation must be as subjective as in any other scientific work (18, p. 402).

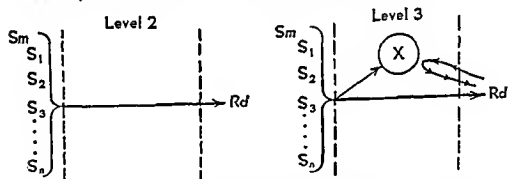
elaboration of the specific operations which fit under the general operational definition of a phenomenon, i.e., the identification of all of the operations producing the phenomenon. *Stimulus variable elaboration* refers to a working out of how different specific operations produce differences in amount of a phenomenon. This, in essence, refers to the search for variables influencing the phenomenon.

*Operational identification* normally takes two steps. First, there is a logical-rational consideration by the scientist (maybe the word "guess" is more appropriate) that a particular set of operations, falling under the general definition of a phenomenon, will produce the phenomenon. Thus, an investigator at some time guessed that falsifying scores (in a downward direction) would produce blocking in a strongly motivated subject and cause a performance difference between experimental (blocked) subjects and control (nonblocked) subjects. If this difference did indeed occur, frustration would be inferred. The second step, of course, was to demonstrate experimentally that a difference between control and experimental subjects occurred as a consequence of difference in treatment. Another experimenter may guess that physical restraint of a small child would constitute blocking, and so on. When two or more operations, differing in detail but fitting the general definition of a phenomenon, have been worked out, operational identification has occurred. Different specific operations may produce differences in amount of the defined phenomenon, and it is apparent that the number of so-called different specific operations will vary depending on how much variation from procedure to procedure is to be called a difference "worth" noting separately. This is entirely an arbitrary matter. Operational identification, therefore, consists essentially in observing that the antecedent conditions of a phenomenon are such as to provoke or give rise to an already established phenomenon. The analytical nature of operational identification and its corresponding contribution to the science is so great, in my opinion, that I shall return to an extended discussion of it in the next chapter.

Now again, however, for S-R definition of phenomena, conceptualization of operational identification may take Level-2 or Level-3 form. Let  $S_{mm}$  stand for the general operations and  $S_{mi}$  with subscript

involved. Other variables will probably influence the extent of the phenomenon and whether or not this is true can be shown by manipulating them in at least two amounts. When such variables are explored at several points we obtain systematic laws of behavior. We not only know that the variable is relevant (it will influence the amount of the phenomenon) but we also know the precise relationship between the two, the degree of precision depending on the number of points explored and the precision or reliability of our response measure.

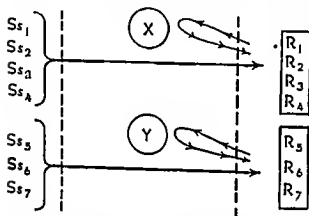
Again, the particular schematic conceptualization of stimulus-variable elaboration will depend on one's bias for Level-2 or Level-3 thinking. The essential idea is that we "put in" the stimulus manipulations defining the phenomenon and then note how other stimulus manipulations (not used in defining) influence the magnitude of the phenomenon. Let us here set off the critical stimulus manipulation as  $Sm$ , and call other stimulus variables  $S_1, S_2$ , etc.



The Level-2 scheme is perfectly straightforward. The phenomenon is said to be such and such a function of each stimulus variable manipulated. Thus, experimental extinction, defined at Level 2, would be said to be related in a certain manner to number of training trials ( $S_1$ ), to ratio of reinforcement during learning ( $S_2$ ), and so on. The Level-3 scheme is somewhat more complicated. Assume we were dealing with *drive*.  $Sm$  is the critical variable, say, deprivation of food in this case. Let us say  $S_1$  is age of animal. If animals of different age show a difference in behavior ( $Rd$ ), we infer that this was caused by a difference in drive. Furthermore, if the relationship between age and behavior is, say, logarithmic, we then say that the relationship between age and drive is logarithmic. In other

Thus, a part of the factor analytic procedure is to make certain guesses about the structures of subject capacities and how these could be "brought out" by certain tests. This activity is indulged in by all scientists; probably no one undertakes an investigation without some thought as to "what will happen." Such thoughts may not be verbalized but that they are almost universally there no one can doubt.

Turning back to factor analysis, let us diagram the situation for these Level-3 type concepts. I shall again speak of static stimuli ( $S_s$ )



and picture the situation after the factor analysis has been completed. To simplify the diagram, I will say that only two factors have been determined (X and Y) and ignore variance unaccounted for. As a result of the correlation among scores on tests 1 through 4,

and of the correlation among scores on tests 5 through 7, and the lack of correlation between the two sets, two factors are inferred, X and Y. Factor analysis fits what I have called operational identification of simple response-defined concepts.

Next, I turn to *stimulus-variable elaboration*, which is largely a refinement of operational identification as far as S-R defined concepts are concerned. If different specific operations produce different amounts of a given phenomenon, the inference is that the operations are allowing different amounts of a particular stimulus variable (environmental, task, or subject) to affect behavior, or that different stimulus variables are involved which influence behavior differentially. At the empirical level, then, the task is one of systematically manipulating potential variables to see what influence they have on the phenomenon. Such manipulation will, of course, occur within the scope of the general operations defining the phenomenon. In the simplest case of the operational definition of a phenomenon, two different amounts of the critical variable are

getting away from the avowed principle of keeping the number of phenomena to be explained at a minimum. Nevertheless, we do not ignore differences for the sake of parsimony; we must note such differences when they occur, keep them conceptualized separately and, when an explanatory system is constructed, account for the differences in the system.

*Are Level-2 concepts "better" than Level-3?* If we limit our considerations solely to the formal operations, Level-2 and Level-3 concepts are identical. The difference comes only in the wording (or implied wording) of the verbal report of the operations. It might seem that Level-3 concepts stray more from the operations than do Level-2 concepts and that this is "bad." The straying comes about, ostensibly, in the naming of a cause for a reliable phenomenon. In the strict sense, with a Level-3 concept nothing more is being said than that a reliable phenomenon has a cause and the name is applied to the hypothetical cause rather than to the phenomenon *per se*. Since we maintain a deterministic position in science, there can be little "wrong" with saying a phenomenon has a cause and in giving this cause (although entirely uncharacterized in the simplest case) a name.

If there is any danger involved it may stem from two sources. The first is that we may sometimes feel we have gone further toward explanation when using Level-3 than when using Level-2 concepts. This is patently not the case, since the operations are the same in both cases. That we may think about the results of the operations differently (thus leading to Level 2 *versus* Level 3) does not add one bit more explanatory validity to one than to the other. The psychologist who defines his concepts at Level 3, irrespective of the amount of elaboration (as discussed earlier) which has taken place, is, at that point, in possession of no more explanatory power than the psychologist who thinks at Level 2. If a psychologist *does* think he has more explanatory power when using Level 3 concepts, if he tends to delude himself that his explanation is further advanced than when he uses Level 2, then there *is* danger in using Level-3 concepts.

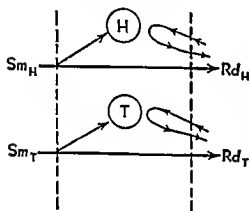
A second danger grows out of the first. I have insisted earlier that when a phenomenon is defined, and when it is elaborated, all we have is a set of relationships. A Level-3 concept does no more than



words, the relationship between a specific stimulus variable and drive is said to be exactly the same as the relationship between this variable and behavior. Thus, the properties of the hypothetical state are directly inferred from and are completely isomorphic to the observed relationship between the manipulable stimulus and behavior.

*Refinements.* As knowledge about a phenomenon grows, changes may take place in the scheme for conceptualization. The first changes may be those indicated above, namely, operational identification and stimulus-variable elaboration. Both of these changes reflect greater inclusiveness. Now, however, changes may occur in the direction of less inclusiveness when data demand it. While several different specific sets of operations may be included

under a general class of operations (all said to be defining a given phenomenon), the lawful relationships between manipulable variables and the phenomenon may vary in some characteristic fashion for different specific sets of operations. Or, the influence of one phenomenon on another may differ as a function of specific sets of operations. Thus, the



influence of thirst drive on a standardized task may differ appreciably from the influence of the hunger drive on the performance on the same task. Degree of interpolated learning may influence retroactive inhibition on a verbal task in a certain way and in a somewhat different way on a motor task. If one is interested in exploring and explaining such differences (and as scientists we would be), the conceptualization would be kept separate. Thus, for hunger and thirst drive at Level 3, we would have two independent sets of relationships as shown in the diagram.

With such a conceptualization, of course, we can have the stimulus-variable elaboration for each drive. If the data demand it, that is, if relationships differ appreciably, such independent conceptualizations must be made. But, we should note that this should be done only when the data demand it, for we are in a serious sense

think the basis for the danger is having a damaging effect on our conceptual thinking.

Now, if you ask me at this point what I mean by explanation I will simply plead that I must postpone further discussion of this issue until later. You may also insist that I am asking for something, namely explanation, which we actually have when we have a set of empirical relationships such as we indeed will have when elaboration of a concept has been far advanced. That is, you may ask me, what is there to explain at the psychological level over and beyond these relationships? At this time I can only say again that the same set of relationships will hold for a Level-2 concept and we go about trying to explain "something" at the psychological level, whereas for the same set of relationships for a concept defined at Level-3 we may not.

#### LEVEL-4 CONCEPTS

*Fundamental characteristics.* Whereas a Level-3 concept is introduced through defining operations, a Level-4 concept is introduced by *postulational* procedure. In its most common form a Level-4 concept is postulated to account for phenomena defined at Level 2. As a very simple case, assume that a psychologist has defined a phenomenon, X, at Level 2. Assume further that he is interested in explaining this phenomenon beyond that given by the critical operation needed to produce the phenomenon and beyond whatever stimulus-variable elaboration may have occurred. The psychologist approaches the explanatory problem initially somewhat as follows: "I will postulate a process (or state or capacity, depending on his predilections and the phenomenon with which he is dealing) Y which is the cause of X." As indicated earlier in the case of a Level-3 concept, the explanatory problem is, in a certain sense, both solved and avoided by definition. In the case of a Level-4 concept, however, the causal process is directly postulated—it is not introduced by defining operations. The mind of the scientist is working differently in the two cases.

*Why Level-4 concepts?* If we postulated a causal process to account for every independently defined Level-2 phenomenon there would obviously be no economy of thought; indeed, there would be considerable redundancy in our conceptual thinking. In the

funnel these relationships through a common term. If the Level-3 concept is a psychological one (as opposed to a physiological one) it has no more locus or substantive existence than does gravity. Yet, it is apparent in the literature that the great bulk of psychologists think of Level-3 as states or processes which have real existence in the body of the organism. Thus, I have continually used such terms in my previous discussion. For most of us it is difficult not to think in such terms. But, the danger is that these Level-3 concepts become spooks, or pixies or elves which have existence in one form or another. Actually, the spooks or pixies, properly sterilized, are probably not such a bad way to think about these processes; most of us disavow belief in the layman's kind of spirits so for us they may perform only as aids to the thought processes. Some might say it would be more dangerous to think of the hypothetical processes in terms of physiological mechanisms which *do* have an aura of reality. Now, I would say immediately that such hypothetical physiological mechanisms *may* have value in directing a search for physiological correlates of the psychological relationships which led the investigator to think in terms of physiological mechanisms. But, if we are searching for explanatory systems at the psychological level, there is a possible danger involved in thinking of Level-3 concepts in terms of physiology. This is because the explanatory attempts at the psychological level may stop at this point.

Therefore, I think the real danger of Level-3 concepts is that they tend to stop explanatory attempts at the psychological level much more than do Level-2 concepts. How many psychologists have attempted to explain *drive* through the use of psychological concepts? How many have attempted to explain *frustration*? Compare the frequency of these attempts for Level-3 concepts with those for Level 2, such as *experimental extinction*, *reminiscence*, and so on. Level-3 concepts impede explanatory attempts at the psychological level; Level-2 concepts invite them, yet both are formally based on exactly the same operations. I think this difference results almost exclusively from the fact that we tend to think that Level-3 concepts imply an existence of a *real* state or process in the organism and, therefore, what more is there to explain at the psychological level. In short, I not only think there is a danger in Level-3 concepts but I

states which enter into accountings made of empirical phenomena. I do not intend to imply that these concepts all have the same formal status; I mean only to imply a commonality is indicated when the investigator has made certain observations and postulates some process to account for these observations.

*A studied intolerance.* Having given a wide range of illustrations of postulated concepts or ideas (some do not have single-word names) I must now dismiss rather abruptly certain kinds of postulated concepts from further discussion. Although I had wished to avoid criticism as much as possible at this stage of the exposition I find right now that I cannot circumvent it altogether. I wish to remove from further consideration those postulated processes which cannot be scientifically defended at our present stage of development. I said earlier that the first purpose of a postulated process is to bring a number of independent phenomena under a minimum number of assumptions. To a greater or lesser extent all of the above illustrations of postulated processes can do this. But, a second and critical purpose of a postulated process or processes is to mediate predictions of new, independently-defined phenomena not used in the original postulation of the process. It can be seen that in order to make testable predictions of these new events, the postulated process must be related in some fashion to stimulus variables (environmental, task, or subject). A new phenomenon depends on a difference in stimulating conditions and unless the postulated process is related in some manner to the stimulating conditions there is no way to make a prediction from the concept idea. Thus, the explanatory idea becomes untestable in the sense of finding out whether it will mediate a prediction or whether one idea predicts better than another. A postulated process not tied to stimulus variables can be assigned all the properties required to explain anything but the validity of these assigned properties can never be assayed experimentally. The soundness of an explanatory idea cannot be evaluated unless the idea is related to at least one stimulus variable. It may have some usefulness as a grouping term in the heuristic sense but it cannot be a part of an explanatory system.

When I say that the postulated process must be tied to a stimulus variable I mean that there must be a stated relationship between changes in the "amount" of the process and variation in the stimulus

initial stages of postulation of a Level-4 concept there is only one purpose which can justify the postulation. The scientist, in postulating the process, brings under a single explanatory idea or principle several independently defined phenomena. That is, phenomena A, B, C, and Y are all said to be manifestations of the single process, X. We thus see that the scientist is, in such instances, adhering to the basic idea of explanation, namely, to account for the greatest number of observations with the fewest number of assumptions. (I shall later discuss the fact that most explanatory attempts of this type actually involve two postulated processes or two independent components of a single process.) So then, the initial purpose of the postulation of a Level-4 concept is to bring several phenomena into a single explanatory orbit.

*Some illustrations of Level-4 concepts.* At the present time I shall place no restrictions on the kind of postulated processes which I will use as illustrations. That is, I shall give these illustrations without any statement as to whether or not they might be considered scientifically valid postulated processes. With all restrictions removed, the number of postulated processes in psychology is legion. Because most of these are well known I shall indicate them by name only.

I suspect that no single idea has been so often reflected in postulated processes as has that of *inhibition*. The essential idea involved is almost always that of a dampening effect on certain types of behavior or a lowering of response potential. The idea of inhibition is old. Sherrington and Pavlov postulated ideas of a *central inhibitory state*, the behavior they observed being largely directed at physiological or neurological mechanisms. We have Hull's *reactive inhibition*, Kohler's *satiation*, Ammons' *temporary work decrement*, Osgood's *reciprocal inhibition*, Glanzer's *stimulus satiation* and even Freud's *ensor* fits the general category. These postulated processes were not, of course, all involved in explanation of the same phenomena, but each was invented as a vehicle to bring several independent phenomena into the same explanatory system.

Now for a few other illustrations not limited to ideas of inhibition. Hull's *anticipatory goal reaction*, Lewin's *tension*, Hebb's *cell assembly*, Krech's *dynamic systems*, Birch and Bitterman's *sensory integration*, Cofer and Foley's *mediated stimulus generalization*, *direction*, as used by Maier, and on and on. These are postulated processes or

and usually do not stand in these isomorphic relationships; rather the characteristics assigned initially are those necessary to mediate (via deduction) the observed behavior. These characteristics are said to vary as some function of the stimulus variables but this same function does not hold between the hypothetical process and behavior. The obvious implication of this is that the hypothetical process or state must in some way modify the input of a given stimulus variable. Furthermore, this modification must come about because either another component of the hypothetical process is related by a different function to the same stimulus variable or a separately postulated process which enters into the particular response under consideration is so related. Thus, measured behavior results from the interaction of these two processes under specified observational conditions. And, of course, there may be more than one stimulus variable which is related to the hypothetical processes and more than one response measure related to the processes in a differential manner. Just what relationships are ascribed depend on the number and nature of phenomena to be incorporated.

The hypothetical relations which are assigned between the stimulus variables and the hypothetical process are, within limits, a product of the scientist's imagination. This may be a somewhat too grandiose term for the scientist's behavior. For, what does he do when he has some reliably established phenomena which he wants to bring under a simple explanatory system? What he does, it appears, is to proceed through a series of trial-and-error, inductive-deductive circles. (If my relatively unsuccessful attempts are any criterion, it is a very agonizing series of inductive-deductive circles.) The scientist has a set of facts; he must assign his hypothetical process (or processes) characteristics which will allow him to deduce these facts. The characteristics he assigns the processes must be related in some fashion to relevant stimulus variables. From one or two facts he may induce certain characteristics that his hypothetical processes must have; he then examines other data to see if they can be deduced from the assigned characteristics. If not, he tries again, this time assigning different properties, and so on, until, if he gets the proper combination, he can "account" for his facts. All the data he has must be represented in the hypothetical processes, either directly or in terms of being deductively generated.

variable. This may be a gross verbal statement, i.e., X increases as some function of the stimulus, or it may be a precise mathematical statement, i.e.,  $X = S_1^2 \times S_2^{-.001}$ . All postulated processes involve a statement of a relationship between the process and behavior if no more than to say that X caused the behavior. When the postulated process or state is not tied to stimulus conditions we are in serious danger of developing spooks or pixies over which we have no scientific control. The spooks multiply, divide, excite, repress or inhibit in a manner dictated by observations of behavior but without reference to stimulus control. Some of the Freudian concepts, such as *libido* and *id* are of this nature, as is also, in my opinion, Hull's *oscillation*. These ideas might be brought under experimental scrutiny by stating what conditions cause them to vary in amount but it is a fact that such statements have not been made. At least in our stage of development we are not equipped to cope with such concepts. Therefore, I am omitting them from further consideration at this time and shall proceed with further characterization of what I will call acceptable Level-4 concepts.

*The growth of Level-4 concepts.* As noted, the scientist initially has reliable phenomena before him when he postulates a Level-4 concept to symbolize processes which he will use in an attempt at explanation. Furthermore, for any given phenomenon, some stimulus-variable elaboration has usually taken place, i.e., the phenomenon is known to vary in certain ways when certain stimulus conditions are manipulated. Now, the postulated process does not remain conceptually amorphous; it is assigned certain characteristics. These characteristics are related (co-vary in some fashion) with the stimulus manipulations and, in turn, the response (behavior) is related to variation in the hypothetical process.

Let me draw a somewhat sharper contrast than probably exists between Level-3 and Level-4 concepts. A Level-3 concept (which is also representative of a hypothetical process) is completely faithful to its defining components, that is, to the stimulus manipulations and to behavior. It transmits perfectly and directly. If a response is shown to be an exponential function of a stimulus variable, the hypothetical process of a Level-3 concept is said to be this same function of the stimulus variable and the response the same function of the hypothetical process. In contrast, Level-4 processes need not

and usually do not stand in these isomorphic relationships; rather the characteristics assigned initially are those necessary to mediate (via deduction) the observed behavior. These characteristics are said to vary as some function of the stimulus variables but this same function does not hold between the hypothetical process and behavior. The obvious implication of this is that the hypothetical process or state must in some way modify the input of a given stimulus variable. Furthermore, this modification must come about because either another component of the hypothetical process is related by a different function to the same stimulus variable or a separately postulated process which enters into the particular response under consideration is so related. Thus, measured behavior results from the interaction of these two processes under specified observational conditions. And, of course, there may be more than one stimulus variable which is related to the hypothetical processes and more than one response measure related to the processes in a differential manner. Just what relationships are ascribed depend on the number and nature of phenomena to be incorporated.

The hypothetical relations which are assigned between the stimulus variables and the hypothetical process are, within limits, a product of the scientist's imagination. This may be a somewhat too grandiose term for the scientist's behavior. For, what does he do when he has some reliably established phenomena which he wants to bring under a simple explanatory system? What he does, it appears, is to proceed through a series of trial-and-error, inductive-deductive circles. (If my relatively unsuccessful attempts are any criterion, it is a very agonizing series of inductive-deductive circles.) The scientist has a set of facts; he must assign his hypothetical process (or processes) characteristics which will allow him to deduce these facts. The characteristics he assigns the processes must be related in some fashion to relevant stimulus variables. From one or two facts he may induce certain characteristics that his hypothetical processes must have; he then examines other data to see if they can be deduced from the assigned characteristics. If not, he tries again, this time assigning different properties, and so on, until, if he gets the proper combination, he can "account" for his facts. All the data he has must be represented in the hypothetical processes, either directly or in terms of being deductively generated.



The initial success of the scientist's efforts depend solely on how many of the available data he can successfully incorporate. But, almost immediately we would also gauge his success on how simply the incorporation takes place. His assigned characteristics cannot be so numerous, and his interactions so complex that for each set of data a special characteristic is required. In the long run the principles used in the accounting must be fewer than the facts for which they account.

Let us suppose now that the scientist has assigned properties to his hypothetical process or processes and that on the basis of a few such properties he can account for an appreciable range of facts. I for one would not minimize such an achievement but at the same time I would insist that he has accomplished only the first stage of scientific explanation. Since properties were assigned to the process on the basis of the facts to be explained they must incorporate those facts; they were assigned so they *would* incorporate them. The scientist must now direct his attention to the predictive capacity of his assigned characteristics. Do these characteristics represent principles with greater generality than those evident in the limited data from which they were derived? What "undiscovered phenomena" can be predicted by the principles? The principles may predict (a) known phenomena not used to induce the principles, or (b) new phenomena (new relationships) which have never been investigated. In both cases the scientist reasons pretty much as follows: "If the characteristics I have assigned these processes (or components of a single process) are valid, then it must be predictable (deducible) that if such and such is done, such and such will happen." If this what "will happen" is independent operationally from those phenomena from which the assigned characteristics were induced in the first place, we have the gratifying experience of a theory predicting, and if the prediction is tested by research and confirmed, we have an even more gratifying experience. But, aside from a consideration of the hedonic state of the investigator we can see that progress is being made in incorporating an expanding body of empirical relationships under a few basic principles. Theoretical notions—the hypothetical processes—constituting a system cannot be closed or sterile. It must have provisions for reaching out and encompassing

new facts in an expanding science. If it does not do this it will sooner or later be replaced by other notions which will do it.

It should be noted that in contrast to Level-4 concepts, Level-3 concepts *per se* have no predictive power over and above the specific relations which led to their definition. Only if the investigator postulates an interaction of a specific kind with another process does the Level-3 concept take on an aura similar to the Level-4 concept.

Having set forth the growth of a Level-4 concept in somewhat abstract, perhaps idealized form, we should turn to a concrete illustration.

*A detailed illustration of a Level-4 concept.* For an illustration I will use the concept of *reactive inhibition*. It is a fairly familiar concept and, therefore, I may hope for some transfer. In discussing the concept I shall also make use of another concept, *excitation*, which interacts with reactive inhibition in determining measured behavior. I will not, however, set forth a detailed account of the history of reactive inhibition. What I wish to show is how the apparent necessity for such a concept arose and how it was said to interact with the excitation process. I will then indicate how these conceptions led to the incorporation of phenomena beyond those which suggested the processes and their characteristics. I shall take the liberty of simplifying certain issues and of trying to reconstruct the thought processes of the scientist. Finally, by way of introducing this illustration, let me say that I am fully aware that it suffers certain shortcomings in being unable to account for some facts and that alternative conceptions have been offered. I am not particularly concerned with the relative worth of the alternative theoretical conceptions but with the nature of this one conception as an illustration of postulational thinking. Reactive inhibition is Hull's term, and my illustration follows Hull, but it should be recognized that there were others before Hull who used essentially the same ideas, notably Pavlov.

Two empirical phenomena led to the idea of an interaction between an excitatory process and an inhibitory process. These phenomena are experimental extinction and spontaneous recovery of conditioned responses. Experimental extinction is the decrement in performance following removal of the unconditioned stimulus, and spontaneous recovery is the increment in performance with the pas-

sage of time following extinction. Now, if we put ourselves in the theorist's position, wanting to unify these two phenomena, what might we suggest? The fact that learning must necessarily take place before extinction indicates that we have to have some process which produces an increment in response strength, i.e., the excitatory component. But, the fact that during extinction the organism responds with less and less magnitude (or frequency, or latency) offers possibilities for choice. In simple form, one could conceive of an actual decrease of the excitatory strength during extinction; or, one might suggest that another process is masking the effect of the excitatory component; or, some combination might be involved. For simplicity and mediating power, the alternative chosen was that of saying that the excitatory strength does not change during extinction; rather, reactive inhibition is built up during extinction thus masking the influence of the excitation. More specifically, it was asserted that every time the organism makes a response a certain amount of reactive inhibition is generated. With such a statement, of course, it means that reactive inhibition is not limited to extinction; it occurs anytime an organism makes a response. Furthermore, the amount of inhibition generated by a response was specified in terms of amount of energy or work required to make the response. Spontaneous recovery would suggest to the theorist (or so it seems in retrospect) that the inhibition which develops must disappear with passage of time. Since the excitation does not change with time, and reactive inhibition does dissipate with time, the passage of time following extinction would leave the excitation component relatively stronger. Thus, on the basis of the two phenomena, the ideas of excitation and reactive inhibition interaction were developed. Now let us see what was done up to this point.

The thought processes of the scientist were inductive initially. There was a certain set of facts available which, in his way of thinking, required the postulation of two interacting processes to account for them. Each process is assigned certain characteristics and a statement is made of how the processes interact. The characteristics assigned each process and the interactive idea are entirely those demanded by the data available. The characteristics assigned are not pulled out of the blue; they are assigned because they will "account" for the obtained results. Other combinations of character-

istics might have served equally well or better, at least initially. When we say, therefore, at this point, that our postulated characteristics account for the observed facts we realize we are being completely circular since we have assigned those characteristics because they *would* account for the facts. I think most would agree that explanatory attempts should not stop at this point.

We must look at the implications of the characteristics assigned; in so doing we arrive at the clear deductive characteristic of our Level-4 concepts. For we ask, if these processes "really" have the characteristics assigned, what new phenomena may be predicted? A little consideration of the postulated characteristics will show that the following may be predicted (to sample a few): distributed practice should give better performance than massed during the acquisition of a conditioned response; massed practice should produce more rapid extinction than distributed practice; differences in rate of spontaneous recovery should occur as a function of number of extinction trials and as a function of amount of work during extinction. Certain predictions can be made concerning alternation behavior. The concepts have been used pretty much as given here in explaining certain rote-learning facts; they have been found useful in accounting for certain facts of motor learning. Thus, a few basic ideas have been shown to be capable of bringing a large number of rather diverse facts together. Let me assert again that at this point I am not interested in the adequacy of this formulation as compared with others; it is the attempt that I am concerned with, since it illustrates well the nature of considerable theorizing in psychology which meets our general criterion of attempting to integrate a large number of empirical relationships by a few basic notions.

Most explanatory attempts, such as the one briefly outlined above, are usually prefaced by an insistence of the scientist that it is a tentative formulation. Failure of prediction of one or more relationships may result in modification or abandonment of the ideas, although the latter is not done lightly. Abandonment of theoretical efforts is not easily accepted, probably because they always have a certain amount of predictive power, even though it be incomplete. Theoretical ideas seem to be lost only when a better set of ideas is available to replace them.

*Are Level-4 concepts defined?* Level-4 concepts are not defined in the sense that Level-1, -2, and -3 concepts are. One does not operationally define a postulated process; one defines a phenomenon, albeit the name may be given to the uncharacterized hypothetical process which is said to cause the phenomenon (Level 3). How, then, do we transmit meaning for Level-4 concepts if not by definition? Scientific meaning is given to Level-4 concepts by relating them to stimulus and response variables. I think Marx's term of *operational validity* is particularly appropriate for this situation (13). The meaning of the concept is given by specifying in what relation it stands to at least one stimulus variable and at least one response variable. These are the minimum requirements demanded for the operational validity of a scientific concept. That is, if the proposed relationships can be put to empirical test directly or indirectly, the concept has operational validity. This testability primarily obtains when the concept is tied to stimulus and response variables. Note, in contrast, that one does not put to empirical test an operational definition; an operational definition, one based on acceptable scientific procedure, is not open to question and as it stands it has no deductive consequences. So, we define Level-4 concepts only in the sense that we characterize them by making statements of their relation to stimulus and response variables, and to other concepts, and that is all we do.

*Are such concepts unique to psychology?* The other sciences have long made use of concepts having the essential characteristics of Level-4 concepts. Level-4 concepts summarize postulated relationships, although in many cases a particular name is not assigned the relationship. But, there are many such names, such as atoms, molecules, and genes which were originally postulated in the manner of a Level-4 concept. There is the postulated characteristic of light energy as corpuscular and an opposing conception of wave-like action. All such notions summarize certain observations and lead, when combined with other concepts, to the prediction of certain relationships which must obtain if the assigned characteristics are valid. Indeed, it seems that postulational behavior of the scientist must essentially be what is usually meant by theorizing.

It should be noted that in the other sciences the postulation of processes or entities akin to Level-4 concepts has sometimes led to observations which confirmed the existence of the process or

entity. Thus, while the gene was originally a postulated entity, later and more refined observations led to the discovery of a structural entity. Some philosophers of science (e.g., 2) and some psychologists (e.g., 11) believe this to be one of the primary functions of such concepts. That is, the characterization of the processes or entities through inferences resulting from experimentally derived relationships may lead investigators to search for structures or processes which were originally built up as convenient but useful fictions. This might not seem on the surface to be an issue of much moment; but, if we ask whether the postulated process should or should not have possibilities of being itself discoverable we find that considerable heat has been generated on the issue with regard to theorizing in psychology. It will be necessary to return to this problem later if we are to succeed in reflecting the temper of contemporary psychology.

I shall not dwell long on these Level-5 concepts. Perhaps ... sense they should be included as a special case of Level-4 concepts. Though they are infrequently used by theorists today and though they share certain characteristics with Level-4 concepts, I should like to keep them separate for the sake of completeness. Essentially these concepts are general *summarizing* concepts; they summarize the interaction of other postulated processes in an explanatory system. Suppose that X and Y are Level-4 concepts which are said to summate to produce the measured response. We might then add a Level-5 concept as a summarizing term and the response is said to be produced by the process indicated by the Level-5 concept. This, of course, would be the simplest possible case and might seem to introduce a redundant concept. But, when several concepts enter into a system several Level-5 concepts may appear and they may help in simplifying the conceptual problem resulting from the interactions of several postulated processes.

I believe the best illustrations of Level-5 concepts are given in Hull's work (7). For example, reaction potential, reactive inhibition, and conditioned inhibition combine to produce effective reaction potential, this latter term therefore being a Level-5 concept. But, in turn, effective reaction potential combines with a hypothetical

oscillatory function (Level 4) to produce momentary effective reaction potential. Thus, we have combinations of combinations at Level 5. I shall spend no more time on these concepts since at the present time they occur in a very small proportion of explanatory attempts.

#### FURTHER COMMENTS ABOUT THE FIVE LEVELS

In reviewing the five levels, let me first repeat some statements made early in the chapter. I am not deceived that the levels are all-inclusive; nor do I feel that it is easy to differentiate the concepts to be included in each. I have found concepts in the literature which I could not discriminate satisfactorily as belonging to only one of the levels. These are cases in which the status of a concept cannot be determined; one cannot tell how it was introduced, for what purpose, and its relationship (if any) to other concepts. And, I have indicated that the same word (summarizing an idea or notion) may be used in different ways (at different levels) by different writers. Yet, there must be some way by which we can sharpen our thinking about concepts used in explanations of behavior and the present is one such attempt. I hope, furthermore, that by using the differences discussed in this chapter and the further differences to follow we may find it possible to arrive at a fair understanding of both the formal status of the concept and possibly some understanding of the thought processes of the scientist as he introduces and uses the concept. But, I see that I may be expecting too much from the written word.

Let me also point out that as thinking develops around a concept its status may change. The levels I have outlined in this chapter are concerned with the initial status of a concept; that is, how it was first introduced by the scientist. There is stability of "placement at a level" only so long as the concept continues to be used in the same manner in which it was first introduced. Inevitably, however, as explanatory attempts grow, there is an elaboration of concepts, especially toward stating of new relationships with other concepts. I think this is especially true with Level-3 concepts. Introduced originally by a definitional procedure they may soon be related to Level-4 concepts and indeed, take on the aura of Level-4 concepts. Their

initial formal status is often obscured by these relationships. Thus, while the Level-3 concept is introduced by definition (not postulation) it may soon be said—be guessed or postulated—to be related to other processes in such and such a manner. For example, *drive*, introduced by definition, may be postulated to interact in such and such a way with associative strength of a habit, or may be said to interact in such and such a manner with other independently defined drives. It is fact that this is fairly common practice in psychology today. Such an elaboration, of course, makes them more than representations of defined phenomena; they become part of an explanatory attempt having the objective of greater inclusiveness. Although I have been somewhat critical of Level-3 concepts (as opposed to Level-2) their saving grace may lie in the apparent fact that the scientist finds it easy to postulate relationships with other processes when the cause is introduced as a part of the defining statement.

*What other distinctions have been made?* I would be misrepresenting the situation if I did not make it plain that many other writers have faced the problems of concept differentiation and have each in his own way made certain distinctions. I shall sample some of the contemporary ones and indicate where they appear to fit into the scheme developed in this chapter.

Taken as a group, concepts at Levels 3, 4, and 5 have been traditionally known as *intervening variables*. A distinction between intervening variables and *hypothetical constructs* made by MacCorquodale and Meehl (11) has suggestions of the distinction made here between Levels 3 and 4. MacKinnon's (12) *phenomenal concept* is similar to Level 3 when operationally elaborated and his *fictional concept* similar to Level 4. If I understand O'Neil (15), his *hypothetical relations* are very nearly the same as Level-2 concepts. His *uncharacterized hypothetical term* might be a Level 3 or a Level 4 at the time the latter first germinates in the scientist's mind. His *characterized hypothetical term* is clearly of the Level-4 type as discussed in this chapter. I should also mention in passing that the diversity of terms is not indigenous to psychologists as can be seen if one turns to the writings of philosophers of science. We all seem to be beset by the plague of individualism in our language. I must indeed apologize for adding more terms to this collection; my only defense (and, common as it is, it is a weak one) is that I found



myself quite incapable of organizing my work around the distinctions which have previously been made.

### REDUCTIVE *VERSUS* NONREDUCTIVE CONCEPTS

As the term is most commonly used in psychology, reductionism means explanation of behavior by means of physiological or neurological concepts. More pointedly, in the context of this chapter, if reductive concepts were introduced at Levels 3, 4, or 5, they would refer to more or less specific neurophysiological mechanisms. And so, presumably, we have another dimension on which concepts may differ, namely, neurophysiological *versus* what I shall call strictly psychological or behavioral concepts. It may be surprising that rather strong opinions prevail among contemporary psychologists on this issue. But they do, and to discover the essentials of current thinking about the place of physiology, if any, among explanatory attempts in psychology I must take a brief foray into this arena.

What is the controversial issue? I suppose it could be said that there are several controversial issues but they all essentially cling around such questions as the following. Should or should not explanatory attempts of psychological events be at the neurophysiological level? Do we have better or "truer" explanation when we use neurophysiological concepts? Should we "require" explanatory concepts to be neurophysiological concepts? Is it possible to have high-order generalizations as axioms in an explanatory system of behavior unless these generalizations are at the neurophysiological level? Why stop at the neurophysiological level; why not the chemical or biochemical or atomic, and so on?

Let me suggest that it is no simple matter to distinguish clearly between psychological and neurophysiological concepts. It is a fact that behavioral events are sometimes used to define and infer physiological phenomena; and so also neurophysiological events are used to infer behavioral phenomena. Therefore, defining operations offer little by way of differentiating disciplines. It seems to me a simple fact that we have difficulty telling just whether a given explanatory concept is physiological or neurophysiological. As a rough means of describing the nature of concepts which may be introduced as

explanatory concepts by psychologists, I will indicate a complex continuum at one end of which are located the strictly psychological or behavioral concepts while at the other end are the strictly neurophysiological concepts.

1. Strictly psychological or behavioral. A concept of this type does not have any neurophysiological implications, is not a term that is ever used at the neurophysiological level and the originator of the term gives no indication that he was thinking in neurophysiological ideas when he introduced the term. I think terms like *frustration*, *reminiscence*, *super-ego*, and *morale* fit this end of the dimension.

2. There are concepts which are neurophysiological "sounding" (i.e., may have been used by neurophysiologists) but no evidence is available that the originator (or subsequent users of the term) was thinking in a neurophysiological manner. *Memory trace* is a concept which has appeared for many years in psychological literature which fits this area on the dimension. So also does *engram*. *Refractory phase* has been occasionally used as an explanatory concept at the psychological level but in a way in which it is quite clear that it is no more than a rough analogy to refractory phase of the neuron.

3. Concepts that may or may not be neurophysiological "sounding," whose usefulness does not depend on neurophysiological facts, but the originator (or subsequent users) made it clear that he was thinking of possible neurophysiological counterparts or physiological mechanisms that might lie behind the behavioral process or state. Many of Hull's concepts are of this nature, e.g., *stimulus trace*. Hull's copious notes leave little doubt that he was continually referring his ideas to the physiological level. Yet, as many have pointed out (almost apologetically) the evaluation of the theoretical ideas fostered by Hull does not depend one bit on references to the physiological data; the system is evaluated at the psychological or behavioral level even though Hull may have found it useful and intriguing to speculate about the neurophysiological counterparts of the relationships at the psychological level.

4. Strictly neurophysiological. In these cases either real or postulated neurophysiological mechanisms are used in explanatory attempts. Thus, theories of vision inevitably make reference to the physiological or perhaps chemical processes. The term has complete neurophysiological implication, it is commonly used at the neuro-

physiological level or an offshoot of this, and the user leaves no doubt that he intends his concept to refer to neurophysiological processes. Kohler's *electrical fields*, Hebb's *phase sequences*, Krech's *dynamic systems* are all concepts which fit at this end of the continuum.

I suspect that the current spate of rather caustic expressions on neurophysiological *versus* psychological theorizing was triggered by MacCorquodale and Meehl (11) in a rather unwitting fashion. In closing an analysis of differences among certain kinds of concepts in psychology these writers suggested that theorists must be more concerned with the reality status of postulated processes or entities. More specifically they said that concepts such as Level-4 concepts should represent entities or processes that "have some probability of being in correspondence with the actual events underlying the behavioral phenomenon, i.e., that the assertions about hypothetical constructs be true" (11, p. 105). What they mean by true is that the construct "should not be manifestly unreal in the sense that it assumes inner events that cannot conceivably occur" (p. 105).

While these writings by MacCorquodale and Meehl do not suggest quite as blunt an issue as I indicated by my introduction, it is clear from the context that these writers feel that in postulating processes the theorist should pay some attention to neurophysiological facts in assigning the properties to postulated processes. Yet, they realize that one might assign properties to a hypothetical process which at a later date would find correspondence—indeed even aid in finding correspondence—in neurophysiology. Their major plea is not to assign properties that are contrary to known neurophysiological facts or that seem highly improbable. These writers would seem to accept theorizing at the psychological level and if knowledge of corresponding neurophysiological mechanisms is lacking, it is quite possible that such theorizing would lead to neurophysiological research in an attempt to discover mechanisms mediating the behavioral process. Thus, it seems to me that all things considered this was a rather subdued plea for psychologists to show greater concern with neurophysiology when theorizing. It remained for Krech to take a completely positive position on this issue. I think Krech's words are worth quoting in order that his position can be clearly understood.

But the moment we introduce hypothetical constructs into our theory building, then the purely psychological approach becomes untenable. I have argued that it is untenable because it makes forever impossible any attempt to approach the study of our hypothetical constructs in any more direct manner than through the examinations of the original stimulus-response correlations. This is so...because the psychological position places hypothetical constructs in a domain which, *by definition*, is forever removed from any direct observation (for that domain...is neither behavioral, experimental or neurological) (10, pp. 287-288).

Where, then, can we place our hypothetical constructs and what *can their nature be?* The answer I have come to, on the basis of all of the above considerations, is a simple one and one which is not at all new. It seems to me that the most fruitful thing to do would be to take the plunge and announce that henceforth our hypothetical constructs (through the use of which we hope to understand all behavior and experience) are to be conceived of as molar neurological events—that and nothing more. Such a step amounts to accepting the universe, and such a step may help us to avoid some of the confusion, esotericisms and circular reasonings that we all have been guilty of in times past. Once having made such a step we would then be in a position to manipulate hypothetical constructs, to have phantasies about the intrinsic attributes of these hypothetical constructs and, on the basis of such hunches, to look for new relationships among the primary data of psychology. And what is most significant, such a step will permit us at least to entertain the hope of eventually being able to study our hypothetical constructs more directly than through guess and hunch (p. 288).

Thus we see that Krech takes quite a dim view of our phantasies at the psychological level of theorizing but would admit, indeed recommend, that such uninhibited phantasizing continue as long as it is directed toward the neurophysiological level. I suspect that the general point of view that we should move rapidly toward theories of behavior based upon neurophysiological concepts has its most reasoned contemporary impetus from the scholarly book by Hebb, published in 1949 (6). Yet, in 1939 Pratt (16) made a very strong plea for a physiological language of explanation. But, because of the *Zeitgeist* or because Pratt said so many other things that aroused controversy so as to obscure this particular stand, the Hebb book stands as a more prominent contemporary landmark. Hebb frankly

sets out to account for certain behavioral phenomena on the basis of neurophysiology of the nervous system. He takes a little bit of neurophysiological fact here, a little there, adds some clearly labelled speculation and arrives at an accounting of some behavioral facts.

Farrell (5) in examining the basic explanatory generalizations in some other sciences notes that they are usually small units, e.g., cells, molecules, atoms, chromosomes. When he then asks himself what comparable units could be obtained at the strictly behavioral level in psychology he reaches an impasse. He then, by logic which is quite unclear to me, concludes, therefore, that psychology must look to neurophysiology for its fundamental generalizations. Experimental research at the behavioral level should continue, he thinks, with the aim of arriving at basic generalizations but with the intent of clearly outlining the behavioral laws for which the neurophysiological mechanisms must account.

And so with samples of one point of view before us, let us look at what is said on the other side. Some quotes from Kessen and Kimble directed specifically at Krech will demonstrate the tenor of this side.

We object to his [Krech's] premise—that psychological concepts *must* be neurological—and to the sort of theorizing to which this conviction leads him. Even more, we oppose the assertion that psychological theory can progress only when we are willing to indulge in neurologistical speculation. In direct contrast, we hold that there is nothing intrinsically more fruitful in physiological theory than in any other kind; further that what Krech calls purely psychological theory is actually in a stronger position insofar as it remains uncluttered by an anachronistic search for "reality" and "true essences" (8, p. 263).

... constructs have no more location than the physicist's concept of force (p. 263).

... theoretical constructs [are] designed to aid in predicting behavior. The extent to which they accomplish this end is a measure of their "value," from which their lack of "neurologicity" subtracts exactly nothing (p. 263).

Our version of the purely psychological psychologist is the scientist who erects his theory and develops his concepts so that the deduced theorems can be confirmed or disproved by observations of behavior. This we demand of him, *and nothing more*. The symbols he uses for theoretical manipulation may have any flavor he likes—neurological,

physical, sociological, aesthetic—but such a psychologist is not required to specify locus or “real” nature in his theory so long as his concepts mediate the prediction of behavior (pp. 263–264).

Adams (1) first takes Krech to task and then later aims his pungent remarks at MacCorquodale and Meehl. Like Kessen and Kimble, Adams points out that the answer to Krech’s petulant inquiry as to where hypothetical constructs exist, is: “In exactly the same ‘physical’ world or Nature as the atom or electron” (p. 68). To Adams, Krech’s question is puzzling but in no sense muzzling. He will not see science hamstrung by any such set of regulations as Krech suggests:

When Krech speaks of “a proper respect for present neurological knowledge and theory” . . . I think he is dead wrong, in spite of his comprehensiveness and important qualifications. The only things to which an inquiry owes respect are its phenomena. The attitude of respect on the part of an empirical science is never appropriate toward existing principles of its own or any other field of inquiry. You break out of the bonds of a doctrine and enlarge it only by *not* having respect for it. We are inherently conservative enough without submitting to such restrictions (p. 69).

And to the MacCorquodale-Meehl suggestion that we should not admit hypothetical constructs which “require the existence of entities and the occurrences of processes which cannot be seriously believed because of other knowledge” (11, p. 106), Adams replies:

. . . when a notion shows a good deal of versatility and seems to be applicable to a variety of phenomena beyond that for which it was designed, it becomes a valued construct, irrespective of the immediate plausibility of the mechanisms it envisages (p. 73, italics omitted).

Finally, with regard to Adams, I should note that he gives Hebb pretty much the same treatment as he has the others. Adams’ position on the issue seems fairly clear.

Bergmann has dissected the assertions of MacCorquodale and Meehl and of Krech and has found them wanting philosophically. He notes that “logically and in principle, physiological reduction is a certainty” (3, p. 442). But that this is true does not in any way eject nonphysiological notions from theories. Relations and proper-

sets out to account for certain behavioral phenomena on the basis of neurophysiology of the nervous system. He takes a little bit of neurophysiological fact here, a little there, adds some clearly labelled speculation and arrives at an accounting of some behavioral facts.

Farrell (5) in examining the basic explanatory generalizations in some other sciences notes that they are usually small units, e.g., cells, molecules, atoms, chromosomes. When he then asks himself what comparable units could be obtained at the strictly behavioral level in psychology he reaches an impasse. He then, by logic which is quite unclear to me, concludes, therefore, that psychology must look to neurophysiology for its fundamental generalizations. Experimental research at the behavioral level should continue, he thinks, with the aim of arriving at basic generalizations but with the intent of clearly outlining the behavioral laws for which the neurophysiological mechanisms must account.

And so with samples of one point of view before us, let us look at what is said on the other side. Some quotes from Kessen and Kimble directed specifically at Krech will demonstrate the tenor of this side.

We object to his [Krech's] premise—that psychological concepts *must* be neurological—and to the sort of theorizing to which this conviction leads him. Even more, we oppose the assertion that psychological theory can progress only when we are willing to indulge in neurological speculation. In direct contrast, we hold that there is nothing intrinsically more fruitful in physiological theory than in any other kind; further that what Krech calls purely psychological theory is actually in a stronger position insofar as it remains uncluttered by an anachronistic search for "reality" and "true essences" (8, p. 263).

. . . constructs have no more location than the physicist's concept of force (p. 263).

. . . theoretical constructs [are] designed to aid in predicting behavior. The extent to which they accomplish this end is a measure of their "value," from which their lack of "neurologicity" subtracts exactly nothing (p. 263).

Our version of the purely psychological psychologist is the scientist who erects his theory and develops his concepts so that the deduced theorems can be confirmed or disproved by observations of behavior. This we demand of him, *and nothing more*. The symbols he uses for theoretical manipulation may have any flavor he likes—neurological,

physical, sociological, aesthetic—but such a psychologist is not required to specify locus or “real” nature in his theory so long as his concepts mediate the prediction of behavior (pp. 263-264).

Adams (1) first takes Krech to task and then later aims his pungent remarks at MacCorquodale and Meehl. Like Kessen and Kimble, Adams points out that the answer to Krech’s petulant inquiry as to where hypothetical constructs exist, is: “In exactly the same ‘physical’ world or Nature as the atom or electron” (p. 68). To Adams, Krech’s question is puzzling but in no sense muzzling. He will not see science hamstrung by any such set of regulations as Krech suggests:

When Krech speaks of “a proper respect for present neurological knowledge and theory” . . . I think he is dead wrong, in spite of his comprehensiveness and important qualifications. The only things to which an inquiry owes respect are its phenomena. The attitude of respect on the part of an empirical science is never appropriate toward existing principles of its own or any other field of inquiry. You break out of the bonds of a doctrine and enlarge it only by *not* having respect for it. We are inherently conservative enough without submitting to such restrictions (p. 69).

And to the MacCorquodale-Meehl suggestion that we should not admit hypothetical constructs which “require the existence of entities and the occurrences of processes which cannot be seriously believed because of other knowledge” (11, p. 106), Adams replies:

. . . when a notion shows a good deal of versatility and seems to be applicable to a variety of phenomena beyond that for which it was designed, it becomes a valued construct, irrespective of the immediate plausibility of the mechanisms it envisages (p. 73, italics omitted).

Finally, with regard to Adams, I should note that he gives Hebb pretty much the same treatment as he has the others. Adams’ position on the issue seems fairly clear.

Bergmann has dissected the assertions of MacCorquodale and Meehl and of Krech and has found them wanting philosophically. He notes that “logically and in principle, physiological reduction is a certainty” (3, p. 442). But that this is true does not in any way eject nonphysiological notions from theories. Relations and proper-



ties (states), the stuff of which theories are made, do not literally occupy space and yet are as real in a scientific and philosophical sense as a nerve or a piece of steel.

I think this is enough on this issue (or is it really an issue?). Others have spoken out at various times and if one wishes to pursue the matter further among the writings of psychologists I would suggest Pratt (16), Marx (14), MacKinnon (12), and Davis (4). I have indicated four areas along a rather complex dimension which might be used as reference points when evaluating the status of a concept. I have also shown that where along this continuum concepts *should* be is a matter of considerable argument. I am afraid that this is one issue at least where I must take a position that someone will describe as vapid eclecticism. But, we are interested in understanding behavior and we should reject nothing which furthers that understanding. Psychology has and will continue to have (in increasing numbers, I believe) physiological concepts in its theories and psychological concepts in its theories. At the present time physiological or mechanical concepts are used almost completely in theories built around the auditory and visual processes. Examine almost any explanatory attempts of the sensory processes and one finds a heavy neuro-physio-mechanical component. In the areas in which empirical knowledge is not so fully developed we may well expect less emphasis on the neurophysiological level. Whether we would be better off in the long run to junk all our psychological constructs and go to the neurophysiological level (as remote as it may seem for many phenomena) is not a matter for group decision. I am sure, and would hope, that we can have theorizing at all levels of discourse. As the sciences slowly unify through overlapping concepts we may see a concomitantly gradual but slow progress toward neuro-physiological-chemical-atomic reduction. But this, if true, represents a later outcome of scientific endeavors; it does not prescribe what the efforts of any scientist shall be at the moment.

## REFERENCES

1. ADAMS, D. K. Learning and explanation. In *The Kentucky symposium*. New York: Wiley, 1954. Quotations reprinted with permission of John Wiley & Sons, Inc.

2. BECK, L. W. Constructions and inferred entities. In H. Feigl & M. Brodbeck (Eds.) *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953.
3. BERGMANN, G. Theoretical psychology. *Ann. Rev. Psychol.*, 1953, 4, 435-458.
4. DAVIS, R. C. Physical psychology. *Psychol. Rev.*, 1953, 60, 7-14.
5. FARRELL, B. A. On the limits of experimental psychology. *Brit. J. Psychol.*, 1955, 46, 165-177.
6. HEBB, D. O. *The organization of behavior*. New York: Wiley, 1949.
7. HULL, C. L. *A behavior system*. New Haven, Yale Univ. Press, 1952.
8. KESSEN, W., & KIMBLE, G. A. "Dynamic systems" and theory construction. *Psychol. Rev.*, 1952, 59, 263-267.
9. KNEALE, W. Induction, explanation, and transcendent hypothesis. In H. Feigl & M. Brodbeck (Eds.) *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953.
10. KRECH, D. Dynamic systems, psychological fields, and hypothetical constructs. *Psychol. Rev.*, 1950, 57, 283-290.
11. MACCORQUODALE, K., & MEEHL, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychol. Rev.*, 1948, 55, 95-107.
12. MACKINNON, D. W. Fact and fancy in personality research. *Amer. Psychol.*, 1953, 8, 138-146.
13. MARX, M. H. The general nature of theory construction. In M. H. Marx (Ed.) *Psychological theory*. New York: Macmillan, 1951.
14. MARX, M. H. Intervening variable or hypothetical construct? *Psychol. Rev.*, 1951, 58, 235-247.
15. O'NEIL, W. M. Hypothetical terms and relations in psychological theorizing. *Brit. J. Psychol.*, 1953, 44, 211-220.
16. PRATT, C. C. *The logic of modern psychology*. New York: Macmillan, 1939.
17. STEVENS, S. S. Psychology: The propaedeutic science. *Phil. Sci.*, 1936, 3, 90-103.
18. THURSTONE, L. L. Psychological implications of factor analysis. *Amer. Psychol.*, 1948, 3, 402-408.

# *The Nature of Some Explanatory Attempts*

## INTRODUCTION

The previous chapter, while intended to examine some differences among concepts and thus help us to understand certain intellectual activities of the scientist, necessarily introduced some preliminary notions of explanation. The present chapter has as its goal a deepening of our comprehension of the explanatory attempts prevalent in psychology. I have made some distinctions among concepts; these distinctions were intended to apply to a concept when it was first introduced by the scientist. I also indicated that there may be immigration of concepts from one level to another as explanatory attempts grow. An explanatory idea is rarely introduced and maintained unaltered. Even in its most strictly abstract form it may grow in assigned properties as it reaches out to encompass a greater range of empirical phenomena than was originally intended. Or, it may shrink in role as other concepts replace some of its functions. So also we find even our most empirical concepts incorporated into explanatory attempts so that they too may acquire attributed properties or characteristics beyond those indicated by the defining operations. It is a by-product of the present chapter to show the nonstatic character of concepts in the hands of theoreticians.

What is explanation? I think I have quite deftly skirted this direct question thus far but this source of ambiguity cannot be tolerated any longer. To answer it as best I can for our purposes, I shall simply make some bald assertions. Within the vision of any one scientist explanation has no end. This is true whether we think of as

yet undiscovered phenomena which must be explained or which will serve to explain, or whether we think in terms of ultimate reduction through the pyramid of the sciences even as we now know them. So, what can I assert? I can assert that scientists engage in activities the outcome of which does two things. First, these activities reduce the number of independent phenomena which require explanation. Secondly, they reduce the number of assumptions necessary to deduce or account for known phenomena. Thus, although we cannot tell when we arrive at the ultimate of explanation, namely, an accounting of all phenomena with the fewest possible assumptions, we can, within the orientation implied by the goal, recognize activities that are commensurate with the orientation. It is my purpose here to sample some of these activities.

One rough dimension which reflects differences among many explanatory attempts is the number of postulated processes (Level 4) involved in the attempt. Perhaps it would be better to say that the dimension represents the ratio between the number of postulated processes and number of empirical phenomena (Levels 2 or 3) which enter into the explanatory attempt. As in all sectioning attempts this one will provide only crude distinctions but it will be satisfactory for my purpose. Therefore, the major part of the chapter will be divided into three sections, namely, *empirical explanation*, *mixed empirical-postulational explanation*, and *postulational explanation*. Within these sections there will also be some other differences which I will point out as we go along. Following these three sections I will discuss some other kinds of explanatory attempts which do not easily fit along the rough dimension noted above. Finally, I find it necessary to make a number of general remarks in order to complete these three chapters on theory or explanation.

### EMPIRICAL EXPLANATION

I suspect that the two words, "empirical explanation," will to many seem contradictory. Yet, within the broad limits I have set for trying to understand explanatory attempts in psychology, many activities may be described by these two words. In the previous chapter, I discussed the idea of *operational identification*. In its most simple and direct form, operational identification is the minimum

activity which I allow as a form of empirical explanation. But, there are less obvious forms of empirical explanation which also must be discussed. As a preview, let me note three possible outcomes of empirical explanation, which, while not independent of each other, ought to be noted separately.

First of all, empirical explanation keeps the number of independent phenomena requiring explanation to a minimum. Secondly, as an outgrowth of persistent attempts at empirical explanation, a given phenomenon may acquire a great deal of generality; that is, it is shown to be a basic behavioral phenomenon in the sense that it occurs under a wide variety of circumstances. Finally, such a generalized phenomenon, along with its relationships to stimulus variables, may become a principle in an explanatory system whereby, along with other principles (empirical or postulated), deductions of other phenomena are possible. The first two outcomes will be apparent in the illustrations of the present section, the third outcome will be illustrated in the following section. Our need at the present, then, is to look at illustrations of the scientist's activities which result in what I am calling empirical explanation.

1. In the "pure" case of operational identification note is taken of the fact that a set of operations used to define one phenomenon are identical or nearly so with a set used to define an already established phenomenon. It is then simply asserted that we are dealing with a single phenomenon. If the comparability of operations is apparent so that the community of scientists will accept the identification, operational reduction is accomplished. Thus, when Zeller (29) noted that his operations, set up to study *repression*, were not critically different from those used to study *motivation*, he concluded that he could not proclaim the experimental isolation of a new phenomenon. Of course, when operations for different research are so similar that they can be said to be studying the same phenomenon by fiat or acclamation, this is usually noted before the researches are carried out. That is, it is noted that the critical variable used in defining a phenomenon is also the critical variable in the research about to be done, and the identification is accomplished. The illustration given in the previous chapter concerning the number of particular ways in which blocking of goal-directed responses may occur to meet the definition of frustration is sufficient and we need not dwell on this

matter. But, I should mention that factor analysis, as a means of keeping the number of independent subject capacities to a minimum, while fitting the idea of empirical explanation, always requires the carrying out of the research to see if the hypothesized relations do indeed exist. I want to turn to instances of empirical explanation where the explanation is less obvious than simple operational identification. Such a situation is more common. Usually what we have is a perceptible difference in operations defining two or more phenomena and it then becomes a matter of research to see if operational identification is justified.

2. Recently an experiment was performed in which subjects, following the learning of a paired-associate list, were presented the responses in the list and were asked to give the stimuli to which the responses had been associated (10). The results showed that considerable amount of such learning had taken place as inferred from the fact that the subjects could recall an appreciable number of the stimuli when presented the responses. Should this phenomenon be kept differentiated from other operations used to define other forms of learning? There is an already established phenomenon called *incidental learning*. The operations for this require that subjects, not instructed to learn anything, do in fact learn something. In the experiment involving the learning of stimuli to responses, subjects were instructed to learn S-R associations but no instructions were given concerning R-S learning, yet the subjects did in fact learn such associations. So, there is a difference in the operations of incidental learning and this R-S learning. The common feature in the operations is that in both instances the subject learned something when not instructed to do so. We might then suggest that essentially we are dealing with the same phenomena.

But, I think that in these situations we should not rush into operational identification. By so doing we may overlook phenomena which do indeed reflect fundamentally different processes of behavior. So what do we do? I think (in the present illustration) the most straightforward manner of reaching a decision is to discover if the variables which influence incidental learning in specified ways will influence R-S learning in comparable ways. In the previous chapter I noted that if the laws relating hunger drive and thirst drive to behavior were appreciably different, we must keep these two

separate. So also, if we find that incidental learning and R-S learning do not respond to variables in the same manner, these also must be kept separate. On the other hand, if behavior does change in much the same way when the same variables are manipulated in the two situations, I think we would arrive at a conclusion that we are essentially dealing with the same processes and that the difference in the two operations is behaviorally irrelevant. Thus, by such procedures, we may avoid talk of two concepts where one is sufficient.

Let us be sure we understand the implications of operational identification as a means of empirical explanation, whether this operational reduction occurs by fiat or as a result of the sort of research described above in the case of incidental learning and R-S learning. The sole initial consequence of operational identification is to keep the number of independent behavioral phenomena to a minimum and in this sense only does it have explanatory value. For example, even if R-S learning is shown to be a simple manifestation of incidental learning, the latter having concept precedence, incidental learning *per se* is not explained. But, the pervasiveness or generality of incidental learning is increased so that it becomes apparent that when it is explained a rather large chunk of behavior will be included under the explanatory system.

3. Operational identification does not always follow the order of events indicated above; that is, it does not always occur by identifying a "new" phenomenon as being in fact produced by operations comparable to an already established phenomenon. There are instances in which one attempts to establish a new phenomenon and then show that it will account for (reproduces the operations of) an already established phenomenon. This peculiar reversal of procedure might seem contrary to my "law" of concept precedence. But, let us get an illustration before us and then evaluate the implications.

A universal phenomenon of serial learning is the bowed serial-position curve. Items at the beginning and end of a list are learned rapidly, those in the middle most slowly. More specifically, the item just past the middle is learned most slowly so that the serial-position curve is nonsymmetrical—it is skewed. Explanations of this skewness have been attempted (e.g., 15) using postulated processes, but in one approach to the problem an attempt was made to account for it by showing it was simply the reflection of another empirical phe-

nomenon but one which had not yet been demonstrated (21). The reasoning was about as follows. The serial position curve makes it possible to say in a factual manner that in learning a serial list the subject acquires items in a forward direction and also in a backward direction. Thus, in a 10-item list, as learning takes place Item 1 will elicit Item 2 before Item 2 will elicit Item 3 and so on for forward learning. Item 9 will elicit Item 10 before 8 will elicit 9 and so on, in the backward direction. The forward and backward learning situations are not operationally comparable. In forward learning a response becomes a stimulus for the next response whereas in backward learning a stimulus becomes a response for the preceding stimulus. Now, after making this analysis, the investigator noted that if backward learning took place more slowly than forward learning the skewness in the serial position curve would be simply a consequence of this fact. However, it would be necessary to show that backward learning did indeed take place more slowly than did forward learning and this had to be demonstrated outside the serial-learning situation. For, if a difference in forward and backward learning is to be used to account for skewness, then the difference must be independent of a situation in which skewness would inevitably occur. This was tested and the results showed that backward learning did take place more slowly than did forward learning.

Having established the difference in forward and backward learning, what does the investigator say? Essentially what he can conclude is that the operations of serial learning allow forward and backward learning to take place; backward learning has been shown to be slower than forward learning. Therefore, the skewness in the serial position curve is explained at the empirical level (is operationally identified with) forward and backward learning. When the difference in forward and backward learning is explained so also will be the skewness in the serial-position curve.

But what about this business of concept precedence which I insisted upon when discussing operational definitions in Chapter 3? Doesn't the present illustration deny concept precedence? Not basically, although I think even if it did we should allow for some flexibility in such arbitrarily trumped-up rules. The idea of concept precedence indicated that when two phenomena had been identified with essentially the same set of operations, precedence is given the



one which has greatest generality. Essentially this means that priority is given the phenomenon which will occur in a situation in which the other could not, by its conceptualized nature, possibly occur. Thus, in the present illustration, "skewness" could not reasonably be expected to occur in a situation that is independent of serial learning. Skewness is tied to serial learning. But, the difference between forward and backward learning is not so tied; it occurs independently of the serial-learning situation, and therefore has precedence. Precedence in situations like this is not achieved by temporal priority of discovery but by generality of the phenomenon involved. I think you would agree that simply because skewness was discovered earlier than was the difference in forward and backward learning we should not insist on saying that skewness caused the difference in forward and backward learning nor even that differences in forward and backward learning are simply a manifestation of skewness.

Although I shall not give any detail, I would like to call attention to the fact that explanations of *spread of effect* based on number biases (e.g., 16) is a case of operational identification in which the empirical demonstration of number biases followed spread of effect.

Let us move along to other forms of empirical explanation which, though in the long run representing a form of operational identification, are somewhat more subtle than those we have examined thus far.

4. We have seen that operational identification as a means of explanation ties directly into our previous material on operational definitions *per se*. And, since operational definitions are intimately related to basic matters of experimental design, we might expect that certain matters of design must inevitably arise when discussing forms of empirical explanation where this is achieved largely by operational identification. One sometimes hears the criticism that we are somewhat overly concerned with details of experimental design in our quest for the purification of relationships and phenomena. We are offered the other alternative of looking for general principles which will subsume the detailed findings; the general principles will supersede any of the minute purifications resulting from close attention to the details of our research operations. None would deny that the search for general principles is a primary goal of science. But I would venture an opinion that one possible way of attaining

this goal is through purification of designs so that fundamental behavioral principles are laid open for inspection. Fundamental behavioral phenomena can only mean phenomena which occur under a wide variety of situations. But, it seems possible to me that this pervasiveness can be masked unless we pay close attention to the fundamental operations involved in our research and do not allow apparent differences in research operations to mask the basic commonality which may underlie several superficially different sets of operations. The following illustration is appropriate.

Approximately 20 years ago, *secondary reward* (or symbolic reward or reinforcement) became a generally accepted phenomenon. Subsequently, experiments began to be reported which, in simplified form, showed that if a rat is given food on every other trial in a simple maze (partial reinforcement) the learning was about as rapid as if food were given after each trial (100 per cent reinforcement). The fact that such a finding was contrary to a general conception of learning is relevant, perhaps, only for understanding the motivation of the scientist who did the work with which I am concerned in this illustration. Rather than accept the findings as indicating a new phenomenon with which the theory must directly cope, this investigator (6) took a careful look at the operations of partial reinforcement experiments and concluded that it was a reasonable guess that the results could be accounted for by the operation of secondary reinforcement. That is to say that there is no new phenomenon involved; the operations for studying partial reinforcement did not eliminate the operation of secondary reinforcement on trials when no primary reward was received. Thus, the secondary reinforcement on every other trial may have resulted in learning nearly as great in amount as would primary reward. If this guess was correct, the removal of secondary rewarding possibilities should result in clear superiority of the 100 per cent reinforcement procedure. Research confirmed this expectation. In short, as far as learning was concerned, the partial reinforcement situation simply allowed for the operation of an already established phenomenon and no new principle of behavior need be assumed. (I am not here concerned with the fact that there now seem clearly to be independent behavioral phenomena associated with partial reinforcement that are not simple manifestations of the operation of secondary reward. I

am interested in the idea here that purifying designs *may* keep the number of independent phenomena to a minimum as well as exhibit the generality of certain phenomena.) There are other instances which also aid in arriving at the conclusion that secondary reward is a very widespread phenomenon. It has now been used essentially as an empirical postulate in certain explanatory systems (e.g., 24) from which, in conjunction with other postulates, other phenomena can be deduced. The generality of the phenomenon of secondary reward, while certainly confirmed by many direct tests, has also been extended by perceptive investigators who have noted that procedures in certain experiments, designed to study different phenomena, were not pure in the sense that they did not exclude the operations used to demonstrate secondary reward. Generality is added by default.

As a matter of fact, some illustrations of errors in design in Chapter 5 are really illustrations of empirical explanation. That is, the scientist simply notes that operations presumed to demonstrate a new phenomenon (again let me remind you that I use phenomenon in a very general way to include even a simple empirical relationship) did not in effect do so because they did not eliminate the possibility that the results represent an already established phenomenon that was allowed to occur. So we see that our designs must keep up with empirical knowledge; the more phenomena we define in an independent fashion, the "purer" must be our subsequent research if we expect to demonstrate unequivocally a new phenomenon. But certainly we must admit that there is a point where the subtlety is so great that to call the confounding an "error" in design is manifestly unfair. I would like to give an illustration where this seems to be the case.

5. In 1939 (4) an experiment was reported that demonstrated what has come to be known as *sensory preconditioning*. In demonstrating this phenomenon two stimuli are paired together over and over. That is, a light and a bell might be presented simultaneously to a dog 200 to 300 times. Then, one of the stimuli is used as a conditioned stimulus in developing a conditioned response, say, leg withdrawal. Then on test trials the other stimulus is presented. If foot withdrawal occurs with greater frequency than appropriate control frequencies, sensory preconditioning is defined. This phe-

nomenon has in general been resistant to attempts to incorporate it into certain learning theories. In 1951 other investigators (28) in studying the operations noted a certain basic similarity between them and other sets of operations used to establish *secondary stimulus generalization* (e.g., 22), this latter phenomenon having been demonstrated in a not too convincing manner but enough so to insure its acceptance as a reliable phenomenon. Now the investigators set up a situation in which a common response would be attached to two stimuli, as is the case in secondary stimulus generalization, but which could also be true for sensory preconditioning. But, in one condition the stimuli were presented simultaneously as in sensory preconditioning and in another separately, as in secondary stimulus generalization. In the second situation, the usual conception of sensory preconditioning would not lead one to expect positive results, i.e., that sensory preconditioning would occur. Yet, both situations would have the common operation of having the same response to both stimuli; this commonality was thought to be the critical part of the operations for both phenomena. If this is the case, both situations should show about the same frequency of response on test trials following the use of one as a conditioned stimulus and the other as the test stimulus, and both should be greater than the controls. This is exactly what happened, and the investigators concluded that sensory preconditioning is just a special case of secondary stimulus generalization; in both instances the critical operation is the making of a common response to two different stimuli. I might mention that this is quite obviously a rather indirect form of operational identification and has just a small assumptive component, namely, that in the sensory preconditioning situation a common response does occur to both stimuli. I doubt if anyone would seriously suggest that this assumption is not warranted.

6. Let me now turn to another case of operational identification in which a behavioral phenomenon may be reduced to reflect completely and faithfully the operation of a simple physiological principle. I say "may" because the reduction process still has to be carried out and may not succeed. But, our main concern is with the logic of the intent.

but the one I will use by way of illustration is foveal contrast. Into one eye two small square fields of light are fed, one called the inducing field and the other the test field. The amount of contrast varies as a function of several factors. For our purposes consider the squares to be joined on one side. As the inducing field gets brighter than the test field, the test field appears to get dimmer. The amount by which it appears to get dimmer is determined by the subject adjusting the brightness of a square focused on the fovea of the other eye. The square, of course, is of the same size as each of the squares on the other fovea. The adjustable brightness square is changed until the subject indicates that its brightness is equal to the brightness of the test field in the other eye.

The empirical explanation which might serve adequately for this phenomenon is *scatter*. Scatter is simply the name for the fact that the vitreous humour of the eye does not transmit light perfectly, slight imperfections allowing the light to scatter somewhat. Thus, if an image with perfectly sharp contours is projected into the eye, because of scatter it will not arrive on the fovea with the same sharpness. That such scattering takes place is a well-established fact. It is believed that brightness contrast may be entirely due to this scatter. To try to effect such an operational identification from a behavioral phenomenon to a physiological phenomenon, two steps are planned. First, the variables which affect brightness scatter will be manipulated again but this time scatter will be measured directly. If the variables should affect scatter in the same way they affect brightness contrast, and if the entire amount of brightness contrast can be shown to be isomorphic with the scatter, it could be concluded that the visual system beyond the fovea is transmitting perfectly that which falls upon it. Secondly, one could separate out subjects who have little scatter and those who have a great deal; two such groups should differ comparably on brightness contrast tests. Again, let me say that whether this will be successful or not is not the issue in our using the illustration here. It is the intent of the investigator that is important for the present discussion. Actually, it would be a rare instance if a behavioral phenomenon could be shown to be completely isomorphic to a physiological phenomenon. Usually, discrepancies develop when it has been attempted to draw parallels between the two levels of description (as we shall see

later) and it may well be that discrepancies will be found in the case of this particular phenomenon.

If the reader is still uncomfortable with the illustrations that I have been giving, that is, uncomfortable in the sense that such illustrations do not square with the usual idea of explanation, I shall now proceed to examples that probably will be found somewhat more consonant with the usual idea of explanation. However, it will be noted in these following illustrations that essentially nothing is being changed; explanation is still being attempted by operational identification.

7. I have indicated that an outcome of persistent operational-identification thinking concerning a phenomenon is that the phenomenon may gather generality. The greater the number of situations in which it is shown to be present, the more it becomes accepted as a general law of behavior. Thus, in the history of a phenomenon we see that we first have a working hypothesis that the phenomenon (not previously shown to have generality) will appear in a new situation where the operations may appear to be superficially different from those used in the original definition. As more and more tests give positive results the generality of the phenomenon becomes accepted. As long as the operations involved in a piece of research include the basic variation defining the phenomenon, certainty that it will appear becomes high regardless of other "distracting" operations involved. Thus, the relationship known as Weber's Law has been confirmed in so many situations that it is expected to be operative in any judgmental situation where variations in magnitude of stimulation occur.

It would be patent misrepresentation were I to leave the impression that this accretion of generality of an empirical phenomenon proceeds smoothly through the succession of researches. Sometimes a phenomenon is found to occur in two or three somewhat different situations and as a consequence of this tonic the investigator becomes somewhat brash, jumping far afield of the essential operations in his speculation concerning the pervasiveness of the phenomenon. This may lead to conceptual clashes if it encroaches upon a domain where other explanatory ideas have taken up squatter's rights. Thus, the attempts to explain certain perceptual phenomena by indicating that they were manifestations of learning was met with resistance by those who held that these phenomena resulted from fundamental

but the one I will use by way of illustration is foveal contrast. Into one eye two small square fields of light are fed, one called the inducing field and the other the test field. The amount of contrast varies as a function of several factors. For our purposes consider the squares to be joined on one side. As the inducing field gets brighter than the test field, the test field appears to get dimmer. The amount by which it appears to get dimmer is determined by the subject adjusting the brightness of a square focused on the fovea of the other eye. The square, of course, is of the same size as each of the squares on the other fovea. The adjustable brightness square is changed until the subject indicates that its brightness is equal to the brightness of the test field in the other eye.

The empirical explanation which might serve adequately for this phenomenon is *scatter*. Scatter is simply the name for the fact that the vitreous humour of the eye does not transmit light perfectly, slight imperfections allowing the light to scatter somewhat. Thus, if an image with perfectly sharp contours is projected into the eye, because of scatter it will not arrive on the fovea with the same sharpness. That such scattering takes place is a well-established fact. It is believed that brightness contrast may be entirely due to this scatter. To try to effect such an operational identification from a behavioral phenomenon to a physiological phenomenon, two steps are planned. First, the variables which affect brightness scatter will be manipulated again but this time scatter will be measured directly. If the variables should affect scatter in the same way they affect brightness contrast, and if the entire amount of brightness contrast can be shown to be isomorphic with the scatter, it could be concluded that the visual system beyond the fovea is transmitting perfectly that which falls upon it. Secondly, one could separate out subjects who have little scatter and those who have a great deal; two such groups should differ comparably on brightness contrast tests. Again, let me say that whether this will be successful or not is not the issue in our using the illustration here. It is the intent of the investigator that is important for the present discussion. Actually, it would be a rare instance if a behavioral phenomenon could be shown to be completely isomorphic to a physiological phenomenon. Usually, discrepancies develop when it has been attempted to draw parallels between the two levels of description (as we shall see

later) and it may well be that discrepancies will be found in the case of this particular phenomenon.

If the reader is still uncomfortable with the illustrations that I have been giving, that is, uncomfortable in the sense that such illustrations do not square with the usual idea of explanation, I shall now proceed to examples that probably will be found somewhat more consonant with the usual idea of explanation. However, it will be noted in these following illustrations that essentially nothing is being changed; explanation is still being attempted by operational identification.

7. I have indicated that an outcome of persistent operational-identification thinking concerning a phenomenon is that the phenomenon may gather generality. The greater the number of situations in which it is shown to be present, the more it becomes accepted as a general law of behavior. Thus, in the history of a phenomenon we see that we first have a working hypothesis that the phenomenon (not previously shown to have generality) will appear in a new situation where the operations may appear to be superficially different from those used in the original definition. As more and more tests give positive results the generality of the phenomenon becomes accepted. As long as the operations involved in a piece of research include the basic variation defining the phenomenon, certainty that it will appear becomes high regardless of other "distracting" operations involved. Thus, the relationship known as Weber's Law has been confirmed in so many situations that it is expected to be operative in any judgmental situation where variations in magnitude of stimulation occur.

It would be patent misrepresentation were I to leave the impression that this accretion of generality of an empirical phenomenon proceeds smoothly through the succession of researches. Sometimes a phenomenon is found to occur in two or three somewhat different situations and as a consequence of this tonic the investigator becomes somewhat brash, jumping far afield of the essential operations in his speculation concerning the pervasiveness of the phenomenon. This may lead to conceptual clashes if it encroaches upon a domain where other explanatory ideas have taken up squatter's rights. Thus, the attempts to explain certain perceptual phenomena by indicating that they were manifestations of learning was met with resistance by those who held that these phenomena resulted from fundamental



properties of sensory organization and were relatively unsullied by learning. In such conflicts, of course, the rigor of research must be substituted for the rancor of words. Empirical phenomena which may at first hold promise of great generality in the investigators' conceptual speculation may eventually be cut back empirically so that the phenomenon will be shown to occur only under a highly restricted set of operations. For example, in my own area of research, distributed practice was thought to be superior to massed practice for learning verbal material of many kinds presented in many ways. Research shows now that this is simply not the case; the phenomenon can be produced only under a highly specific set of conditions. The history of science shows many phenomena which in the initial stages of work upon them gave little indication or promise of achieving the great generality which they later attained.

It seems to me that in the stage of science where there is emphasis on empirical growth (such as I judge psychology to be in at the present time) we may expect many of these attempts at empirical-extension to break out. The scientist seeks for general laws in attempts to avoid being faced continually with pesky isolated sets of data. If, through operational identification, these facts can be shown to be manifestations of a few basic phenomena the science advances rapidly. One can note these tentative probings in several areas in our literature at the present time. The recent emphasis on motivational factors in perception is one illustration. As another, one of my colleagues (8) systematically compared the influence of certain variables on some motor learning phenomena and the influence of these same variables on certain perceptual phenomena. The striking comparability has led him to suggest that we are not dealing with disparate functions in the two fields. Helson's *adaptation level* (13), which is basically an empirical phenomenon, has been demonstrated in several judgmental situations and it is tentatively suggested that it will occur in a much wider range of situations. Berg (3) has suggested that *response sets*, shown in a number of situations, may also operate in personality and interest inventories and that the scores on these tests may reflect largely differential response sets which are independent of the content of the test item. I suspect that when the implication of this attempt at empirical extension is fully realized we may expect words to fly.

So we see that throughout our science, attempts at operational identification are being carried on constantly. Their consequence is to keep the number of independent phenomena down to a minimum and to determine the degree of generality of these phenomena. They result from what may itself be a fundamental principle of behavior, namely, to think of new things in terms of things about which we already know.

#### MIXED EMPIRICAL-POSTULATIONAL EXPLANATION

In empirical explanation as discussed in the previous section neither postulated processes nor hypothetical properties were involved. If there is a hypothetical component involved in empirical explanation it is usually no more than a working hypothesis that "this" phenomenon will occur as a result of this set of operations. Usually there is no deductive component directly involved in the operational identification. I have tried to show that operational identification is a basic scientific activity, not mere pedantry, and that these simple working hypotheses that often guide operational reduction may have startling implications when they trespass on supposedly well-conceptualized areas of behavior. In the present section, while we fully realize the futility of trying to hold to clear-cut categories in material of this sort, we propose to proceed to a discussion of explanatory attempts which, like operational reduction, start with the working hypothesis which directs operational identification. But, in addition, since simple operational identification will not incorporate the phenomena under consideration, postulational steps are necessary.

I suspect that empirical-postulational explanatory attempts will have much more of the aura of "true" theory than did empirical explanation. Furthermore, explanatory attempts by empirical-postulational techniques are widespread in psychology and I will not be able to give this type of explanation adequate representation in the space I feel I can allot to it. By their nature, they are simply more complicated than empirical explanations and even though I cut corners in giving illustrations they must take considerable space. We have books intended to incorporate broad areas of behavior into an explanatory system which uses this mixture of operational reduction

properties of sensory organization and were relatively unsullied by learning. In such conflicts, of course, the rigor of research must be substituted for the rancor of words. Empirical phenomena which may at first hold promise of great generality in the investigators' conceptual speculation may eventually be cut back empirically so that the phenomenon will be shown to occur only under a highly restricted set of operations. For example, in my own area of research, distributed practice was thought to be superior to massed practice for learning verbal material of many kinds presented in many ways. Research shows now that this is simply not the case; the phenomenon can be produced only under a highly specific set of conditions. The history of science shows many phenomena which in the initial stages of work upon them gave little indication or promise of achieving the great generality which they later attained.

It seems to me that in the stage of science where there is emphasis on empirical growth (such as I judge psychology to be in at the present time) we may expect many of these attempts at empirical-extension to break out. The scientist seeks for general laws in attempts to avoid being faced continually with pesky isolated sets of data. If, through operational identification, these facts can be shown to be manifestations of a few basic phenomena the science advances rapidly. One can note these tentative probings in several areas in our literature at the present time. The recent emphasis on motivational factors in perception is one illustration. As another, one of my colleagues (8) systematically compared the influence of certain variables on some motor learning phenomena and the influence of these same variables on certain perceptual phenomena. The striking comparability has led him to suggest that we are not dealing with disparate functions in the two fields. Helson's *adaptation level* (13), which is basically an empirical phenomenon, has been demonstrated in several judgmental situations and it is tentatively suggested that it will occur in a much wider range of situations. Berg (3) has suggested that *response sets*, shown in a number of situations, may also operate in personality and interest inventories and that the scores on these tests may reflect largely differential response sets which are independent of the content of the test item. I suspect that when the implication of this attempt at empirical extension is fully realized we may expect words to fly.

So we see that throughout our science, attempts at operational identification are being carried on constantly. Their consequence is to keep the number of independent phenomena down to a minimum and to determine the degree of generality of these phenomena. They result from what may itself be a fundamental principle of behavior, namely, to think of new things in terms of things about which we already know.

#### MIXED EMPIRICAL-POSTULATIONAL EXPLANATION

In empirical explanation as discussed in the previous section neither postulated processes nor hypothetical properties were involved. If there is a hypothetical component involved in empirical explanation it is usually no more than a working hypothesis that "this" phenomenon will occur as a result of this set of operations. Usually there is no deductive component directly involved in the operational identification. I have tried to show that operational identification is a basic scientific activity, not mere pedantry, and that these simple working hypotheses that often guide operational reduction may have startling implications when they trespass on supposedly well-conceptualized areas of behavior. In the present section, while we fully realize the futility of trying to hold to clear-cut categories in material of this sort, we propose to proceed to a discussion of explanatory attempts which, like operational reduction, start with the working hypothesis which directs operational identification. But, in addition, since simple operational identification will not incorporate the phenomena under consideration, postulational steps are necessary.

I suspect that empirical-postulational explanatory attempts will have much more of the aura of "true" theory than did empirical explanation. Furthermore, explanatory attempts by empirical-postulational techniques are widespread in psychology and I will not be able to give this type of explanation adequate representation in the space I feel I can allot to it. By their nature, they are simply more complicated than empirical explanations and even though I cut corners in giving illustrations they must take considerable space. We have books intended to incorporate broad areas of behavior into an explanatory system which uses this mixture of operational reduction

and postulation. For example, *Personality and Psychotherapy* by Miller and Dollard (7) assumes the reliability of a wide variety of clinical phenomena, then by a deft mixture of operational identification and postulated processes and relationships attempts to account for these phenomena. I cannot, of course, review such a system and must stick to my policy of examining miniature explanatory attempts.

1. In the general area of learning and retention, I suppose that no empirical phenomenon enters into more explanatory systems than does stimulus generalization. The use of this phenomenon in a miniature explanatory system may be illustrated by Spence's classic account of certain facts of discrimination learning in animals (23). More specifically, it was an accounting of discrimination-learning facts which suggested that animals learn relations among stimuli, thus leading to what is known as transposition behavior. If an animal learns to approach a 5-unit stimulus and avoid a 10-unit stimulus, then when a 10-unit stimulus and a 15-unit stimulus are presented together, transposition is said to occur if the animal chooses the 10-unit stimulus. Such behavior led to the belief by some that the animal learned to "choose the smaller or weaker of two stimuli" so that when the 10- and 15-unit stimuli were presented together it chose the smaller. Spence's approach was to show that this interpretation is not necessarily required.

First, the phenomenon of stimulus generalization was assumed to be operating in such learning situations. Since stimulus generalization had been independently demonstrated in many learning situations the assumption is hardly open to question. When an animal is rewarded for making an approach response to a stimulus (the correct response) a gradient of stimulus generalization can be shown empirically to exist. There was fragmentary evidence that a similar gradient might exist around the negative stimulus, which means that not only did the animal learn to avoid the particular stimulus which was not rewarded but also other stimuli similar to it, with the tendency to avoid being less and less as stimuli became less and less similar to the negative training stimulus. This is the entire empirical content of the theory. But in addition, certain characteristics were postulated for these phenomena. (a) Particular shapes for the positive and negative gradients of stimulus generalization

were postulated; (b) a postulated interaction that whenever the two gradients overlap the response tendencies summate algebraically; (c) the animal will always respond to the stimulus with the greatest net positive habit strength. With this system Spence could predict transposition behavior and could also predict in what situations transposition behavior would fail to occur. This latter prediction would not be expected by one who used the relational approach; thus, supposedly, a test of two opposing conceptions was possible.

Now what did Spence do in constructing this theory? He took an empirical phenomenon, stimulus generalization, which I would call a Level-2 concept. But, in assigning the hypothetical properties (particular shapes to the generalization gradient and interaction) he is using what I have called Level-4 thinking. Whether we should now call stimulus generalization a Level-4 concept or not, or whether one even wants to think in terms of levels is relatively unimportant as long as we see the nature of the scientist's thinking which is involved. I have maintained that we do not work with static concepts; this is an illustration of that fact. I would also like to note again, as I did in the previous chapter, that explanatory attempts involving postulated processes always involve two or more processes and a postulated interaction between them.

2. We have already seen how Hebb's (12) theorizing leads him to the neurophysiological level. It is my belief, however, that whether the scientist "lowers" himself to the neurophysiological level or whether he "raises" himself to the abstract level, the fundamental characteristics of his thought processes do not differ. Research resulting from the different practices may differ in that the neurophysiological reductionist is making assertions about the nature of neurophysiological processes and may, therefore, attempt to determine independently the validity of these assertions by neurophysiological research. But, it need not be so. However, let us take a look at a fragment of Hebb's theorizing and hope we won't be doing him a serious injustice as we illustrate the empirical-postulational approach. For simplicity, I shall state that the primary behavioral phenomenon which Hebb is trying to explain in this system within a system is memory. The summary name for his essential explanatory mechanism is the *cell assembly*. What he tries to do is

explain how memories are established. To do this he believes he must have a mechanism for a temporary reverberation of the neural trace resulting from a perception. Also, there must be some growth change during the reverberation which effects a permanent modification (memory). What does he do to accomplish this? First, he gathers together all the neurophysiological facts he can which might be relevant, i.e., the fact that neurons have certain structures, that there are synaptic junctions between cells, that one cell may fire another, and so on. He also shows that evidence is fairly clear on the fact that there is a reverberatory trace occurring among simple neuron circuits under certain conditions. But, also looking at the facts on refractory phase, and the fact that a simple circuit of neurons could not possibly reverberate long enough to establish permanent change, he makes certain assumptions or postulates. The essential postulate is that if neuron circuits converged in a certain way, and if alternative pathways of reverberation were possible as a consequence, permanent change could be set up to account for memory and certain allied phenomena. This is a simplified version and perhaps not exactly accurate, but it is a close enough approximation to show the nature of Hebb's merging of neurophysiological fact and his fiction (postulated characteristics) about them to account for certain behavioral phenomena. Possibly it is not quite accurate to say that this accounts for the behavioral phenomena; it is perhaps more accurate to say that the proposed neurophysiological events underlie the behavioral phenomena.

Considering Hebb's and Spence's illustrations of theorizing together, I think there are three points that I would like to make partly by way of reminders of previous discussion. First, why did Spence and why did Hebb postulate the particular processes they did? The reason is, as we have pointed out in the previous chapter, that postulated characteristics are assigned in such a way as to account for the behavioral phenomena. In this sense the reasoning reflects the inductive-deductive circles I have talked about earlier. We must not allow ourselves to believe that in the initial formulation of such an explanatory attempt it is anything but an *ad-hoc* formulation; it is built to accommodate certain facts and therefore must accommodate them. Only when it is shown that it will incorporate facts not used in its construction, and which are operationally

independent from those so used, does the explanation lose its "*ad hocness*" and become a predictive system. And again, I must insist that to say an explanation is *ad hoc* is not to damn either it or the scientific activity which led to it. For, to find the proper combination of postulated characteristics (in these cases used in conjunction with certain empirical relationships) that will generate the known facts is not an easy job as long as the scientist adheres to the principle that the number of postulated processes must be appreciably fewer than the number of operationally independent phenomena for which these processes are to account.

Secondly, it might seem at first in looking at Spence's approach *versus* Hebb's approach that Hebb's postulational attempts are much more circumscribed by the situation than is Spence's. That is, it might be thought that Hebb must work at the postulational level within a very narrow range of possibilities because he is so restricted by neurophysiological facts against which his postulates must not run contrary. If this be true, however, it must be clear that this is not a consequence of reductionism *versus* nonreductionism. In the particular kind of explanation which I am considering (combined empirical and postulational) the scientist is bound only by the known empirical phenomena (and their relationships with known variables) which are used as the empirical component of the explanatory attempt. Thus, while Spence postulated the specific form of the generalization gradient, it did not contradict the essential but general fact that the gradient falls as some function of similarity to the training stimulus. Likewise, Hebb did not postulate any characteristics which contradicted any known characteristics of the nervous impulse. Thus, how much one may be initially restricted in his explanatory attempts depends upon the amount of empirical content he starts with. We might have instances in which the empirical content was low, hence little restriction is imposed on the nature of the postulated processes. On the other hand, if we bring a great deal of empirical content as a base for an explanatory system we may (but do not necessarily) restrict the diversity of the postulational attempts. One might contrast Spence's theory of discrimination learning where the empirical content is low and postulated content is high, with his theory of delay of reward learning where the empirical content is relatively high and the postulated



relatively low (24). I do not know, however, whether or not these two parts of an explanatory system are always reciprocals; the only point I wish to make is that it does not seem to me that the reductionist is any more restricted in his fantasies than is the nonreductionist. Any differences in restriction that occur in empirical-postulational explanation probably result from differences in extensiveness of empirical phenomena and laws about them that are put into the explanatory system in the first place.

Third, I should note briefly how the empirical-postulational explanatory attempts may lead to two different ways of verification. First, we may in many cases test directly the postulated characteristics assigned the empirical phenomena of the system. Thus, Spence postulated a specific form of the generalization gradient and one could test this directly. If such a test showed that the postulated shape does not exist in fact then the system would have to be altered to encompass this necessary change. So also may neurophysiological evidence confirm or deny some of Hebb's postulated processes. Secondly, one may empirically test deductions which stem from the system, and I think we would agree that this is less direct (although no more or less valid or critical) than the direct tests of assumptions. Either method leads to clarification and evaluation of explanatory attempts. Indeed, if in the neurophysiological form of theorizing all assumptions are tested and confirmed then the explanatory system becomes no more than an elaborate form of strict empirical explanation arrived at in a different manner than was discussed for such forms of explanation. However, the same degree of certainty can be attained by positive tests of postulated characteristics at the abstract level although I suspect that for equal degrees of certainty a greater number of tests is required when dealing with abstract processes. But let me now turn back to further illustrations of empirical-postulational explanation.

3. As I have suggested earlier, explanatory attempts dealing with behavior associated with sensory processes have rather consistently turned to neurophysiological mechanisms. Most of these explanations must necessarily take on the combined empirical-postulational approach. The empirical content usually refers to neurophysiological (or chemical, or electrical, or mechanical) facts and the postulational procedures are used to fill in the gaps not covered by these facts.

For example, certain theories of hearing, in an effort to account for phenomenal pitch, start with the fact that the basilar membrane exists in a closed "container" surrounded by fluid. This fluid, following the physical laws of fluid in a closed system, will take certain wave patterns when the flexible walls of the cochlea are flexed by pressure changes occurring as a result of auditory stimulation. The postulation comes in suggesting how the wave patterns may stimulate the cells of the basilar membrane and how subsequent nerve impulses could mediate phenomenal pitch. The range of postulation is also restricted by certain known facts such as rate of firing of certain nerve cells, pattern of firings in the nerve trunk, and so on.

Consider the visual modality. It is a well-established fact that in the rods of the retina there is a chemical substance, rhodopsin, which changes (bleaches) when exposed to light. Other facts, which I shall not detail, show almost beyond doubt that this substance is involved critically in instigating impulses in the optic nerve, but just how this is done (as well as filling in other gaps) becomes a matter of postulation. In general, I think it is a fair evaluation to say that neurophysiological facts enter strongly into most explanations dealing with sensory phenomena and that postulational efforts are directed toward filling in gaps in this neurophysiological knowledge.

### POSTULATIONAL EXPLANATIONS

In the previous section we have seen how empirical laws and postulated "laws" (law-like statements of hypothetical processes) may be used to mediate empirical phenomena. With pure postulational explanations the basic relationships from which deductions are made are all postulated. There are a few pure cases of such explanatory systems in psychology. From the point of view of the philosopher of science who may feel strongly about the elegance of explanatory systems, I suspect these pure postulational systems represent works of beauty. The most notable illustration in psychology of the pure postulational approach is the *Mathematico-Deductive Theory of Rote Learning*, by Hull *et al* (15). This work contains 18 postulates not one of which is a direct statement of an empirically-determined behavioral law. The postulates were arrived at because deductions from them would account for a wide range of

phenomena in the area of rote learning. Subsequent work by Hull uses both empirical and postulated laws in the systems. For our purposes here, we will examine a miniature system dealing with a relatively restricted area of animal behavior.

In the area of animal learning there is a well-established phenomenon of alternation behavior. This behavior is most clearly seen in a simple T-maze. If food is placed in both arms of the maze and the animal given a series of massed trials, it will tend to alternate between the two arms on successive trials. Several variables influence the extent or amount of alternation. Different theoretical accountings have been offered for these facts; the one I wish to discuss is a near pure case of postulational explanation. This theory, developed by Glanzer (11), is presented as one postulate with several parts. However, I shall outline it here in different form and omit certain aspects which are not necessary for our purpose.

*Postulate 1.* Each moment an organism perceives a stimulus-object there develops a quantity of stimulus satiation to the object. Stimulus satiation reduces the organism's tendency to make any response to that object.

*Postulate 2.* The same amount of stimulus satiation to the object develops in each successive moment. The total amount developed, therefore, is an increasing linear function of time. There is a loss of part of each quantity of stimulus satiation in each successive moment. The amount of stimulus satiation remaining from each quantity is a decreasing negative exponential function of time.

*Postulate 3.* Stimulus satiation developed to an object will be generalized to other stimulus objects as a direct function of similarity of the objects.

*Postulate 4.* Various quantities of stimulus satiation combine additively.

Now, let us see what we have. We are trying to give an accounting of alternation phenomena, these phenomena usually being defined as Level-2 concepts. The single hypothetical process which is involved is *stimulus satiation*; this is clearly a Level-4 concept. The characteristics assigned stimulus satiation by postulation are those characteristics needed to account for the fact of alternation behavior and the influence of certain stimulus variables on its amount. Postulate 1 as stated assumes that an organism may develop

a tendency to respond to an object but makes no provisions for how this tendency has developed. We might suspect that this tendency to respond developed largely from learning; however, the first postulate says only that stimulus satiation develops to an object as a consequence of perceiving it and irrespective of how the tendency arose to respond to the object, stimulus satiation will reduce that tendency.

Postulate 2 is a quantitative statement of the development and dissipation of stimulus satiation. The postulate says that if an animal stares (to put it crudely) at an object, stimulus satiation develops linearly but at the same time dissipates as a negative exponential function of time. Obviously, to determine the net satiation at any moment the accretion and dissipation functions must somehow be summed. This Glanzer does, so that the amount of stimulus satiation effective as a function of perceiving time is an increasing negative exponential function of time; this statement, in effect, becomes a first deduction from a consideration of the properties assigned the accretion and dissipation of the satiation.

Postulate 3 comes near to being an empirical postulate, but it actually isn't. It is an empirical fact that response tendencies do generalize as a function of similarity, but Glanzer is saying that this hypothetical state also generalizes, and one cannot have a true empirical law of a hypothetical process. Postulate 4 is needed, among other reasons, so that stimulus satiation developed to two or more similar stimulus objects can be functionally unified.

Why was this process of stimulus satiation postulated and why were its particular characteristics postulated? To repeat what I have said on several occasions earlier, this was done because by so doing they will account for certain facts; that is, the facts were used to induce the process and its characteristics. But, with these characteristics available, certain new phenomena are predicted and it is thus possible to evaluate somewhat the effectiveness of the formulation. The system should, of course, predict the phenomena on which it was built; but will it predict something new? Let us look first at the basic phenomena. If the animal on one trial takes one arm of a T-maze, a certain amount of stimulus satiation will be built up so that on an immediately succeeding trial the animal will very likely take the other arm since no stimulus satiation (or at least less) exists

for it. It can also be seen that as time between trials increases, alternation behavior will decrease, for the satiation will dissipate in the interval between trials. A number of other established facts also follow. Then, Glanzer lists a number of predictions which stem from the theory which have not as yet been tested. He further explores the implications of the postulates for situations with more than two alternatives. And he also suggests how other phenomena, not immediately seen to be related to alternation behavior, may be deduced from the system. For example, *exploratory behavior*, usually thought of as a special kind of drive, may be deduced from the postulates. Finally, he makes tentative suggestions concerning certain human behavior which might be subsumed under such a set of postulates.

Now, whether one likes the approach Glanzer has used or not, and whether or not there are certain vaguenesses (e.g., how do we tell when perceiving is occurring?), I think we must admit that Glanzer has in general taken into account certain "rules" of theory construction which are often suggested. First, he has relatively few assumptions; he has shown how these assumptions may account for a number of facts about alternation behavior; he has indicated how implications of the theory may be tested by performing new sets of operations, and he explores the implications of the theory far beyond the facts associated with simple alternation behavior. Whether it is as satisfactory as other attempts in accounting for the same behavior is up to the experts in the particular field to decide.

I shall give no further illustrations of the strict postulational approach. Our discussion of Level-4 concepts gave us considerable by way of illustrating the thinking which goes into these kinds of explanatory systems and furthermore, there is no essential difference between the combined empirical-postulational approach and the postulational once the postulates are brought together. It should be noted, however, that in the sense of making direct tests of postulates, as we can when an empirical postulate is involved, it is not often possible when the approach is purely postulational. Many tests must be made in terms of implications of the interaction of the processes and their characteristics.

Let me review what I have done thus far in this chapter. First I presented illustrations of what I have called empirical explanation,

from simple operational identification to more subtle forms whereby research was involved to see if two or more phenomena, originally believed to be independent, could be fitted into a single basic set of operations. I then gave illustrations of some explanatory attempts which combine empirical phenomena or laws and postulated processes as axioms or principles from which phenomena are deduced. The third form of explanation was the pure postulational. In the previous chapter various types of concepts used throughout psychology were discussed, certain of these concepts being definitely explanatory ones. When we put together the material of this previous chapter with that of the present chapter have we exhausted the theoretical behavior shown by psychologists? Of course not, although I would venture a statistical opinion that in one way or another the discussion has included the bulk of such explanatory behavior which is clearly oriented toward research findings. Nevertheless, I now want to proceed with consideration of other explanatory approaches which do not fit well into my previous sections. Finally, I will discuss other issues which are related to the general problem of explanation in psychology.

### MODELS

In light of current trends in psychology it is almost obligatory that I say something about models as explanatory devices. As might be expected, this type of explanation is by no means unrelated to the types we have previously discussed. I say this "type" of explanation as if there was a particular definable way in which the term is used by all writers. By now I think we know better than to expect such good fortune. While it is my understanding that the term does have a fairly specific meaning in the physical sciences, it does not in psychology. I am not going to state a specific definition but rather will discuss a number of different scientific activities which have at one time or another been referred to as making use of models.

If I may be allowed a somewhat specious statement for purposes of making what seems to me a fairly relevant point, it is that the use of models is simply another means by which we attempt to understand the unknown through the use of the known. I say another means because if we examine the explanatory attempts already

discussed we note this fundamental theme. Operational identification (empirical explanation) is clearly this sort of thinking; and, when postulated processes enter into explanatory systems the characteristics of these processes almost inevitably are similar to something about which we already know (20). It seems intellectually compatible, almost to the point of necessity, that we think of new phenomena in terms of events and relationships and characteristics about which we already know. Yet the seeming inevitableness of this mode of thought should not make us complacent, for without doubt many of the great strokes of explanatory genius have come about because the scientist did break through this intellectual restriction and allowed his imagination to focus on implications of relationships foreign to those about which he had previously thought.

For our purposes, the term "model" may be said to be introduced in two different contexts. Whether or not these two contexts are completely distinct is of little consequence; at least an examination of these contexts will give us some insight into activities of research psychologists which we have not hitherto discussed fully.

*Research models.* Research in a relatively new area of investigation is seldom undertaken without some conceptual scheme in mind. That is, it is seldom undertaken without some preconception as to the nature of the phenomena and perhaps the processes lying behind them. These predilections are usually lightly held but they do afford the initial working hypotheses, i.e., what variables to investigate initially. If one studied the personal history of the particular scientist involved one could probably determine the source of these orienting attitudes, as they have been recently called (9). Let me give you an illustration of what I mean by these research models. In the last few years there has been a rather marked growth in interest in studying the thought processes experimentally. Several investigators have offered research models which they believe might be useful in the initial attacks on the area. Thus, Bartlett (2) views thinking as having a counterpart in motor learning, and the research he would undertake would be directed initially by this conception. Kendler (17) believes certain phenomena of simple conditioning will be evident in problem-solving behavior. From such conceptions certain variables are suggested as being important; hence the initial

investigations are directed toward determining the influence of these variables.

In the above illustrations the research model was derived from another area of research in psychology. But models may also come from other disciplines, both empirical disciplines and the formal discipline of mathematics or statistics. For example, some (26, 27) have suggested that thinking may be fruitfully conceived of as a probability matter and within certain limits will follow the laws of probability. Such conceptions indicate variables which may be highly relevant. Other investigators have viewed concept formation as a matter of information processing (1, 14) and, therefore, the factors which are important in information processing may be pertinent to concept learning.

I shall note later how these conceptual schemes may be introduced with a somewhat different purpose in mind than I have indicated. For the moment, however, I am concerned with these research models that are set up with the primary purpose of getting research "hooks" into an area in some planned manner. What are the various outcomes of the use of such research models?

One possible outcome of the initial use of a research model is that the variables suggested by the model are not highly relevant. Two or three studies may show that the model has little or no counterpart in the particular behavior being investigated; the model isn't sterile, it just isn't very relevant. Another possibility is that the model does suggest two or three very relevant variables but beyond this has little to offer. Another outcome is that the model may suggest many variables and research shows these to be very relevant. Even a fertile model such as this may subsequently be abandoned, for as the system of relationships becomes developed in the new area of research they may become conceptually organized within themselves; the original model has served its purpose and the newly organized conceptual system will itself suggest the additional research attacks.

Suppose we have used the phenomena and associated relationships in one area of psychology as a model for a new area of research also in psychology. And suppose that the model serves well in the sense that in the new area of research the model phenomena are found to be present and the laws relating them to the stimulus variables are found to be quite similar. If this happens, we essentially



inculcate the new area of research under the old and if an explanatory system has been developed for the old it will serve for the new. I think it is evident that when such identification is found to be even partially complete, we are dealing with a complex case of empirical explanation as I discussed it earlier in the chapter. For, the result is to keep the number of independent phenomena to a minimum by operational identification. So, the research model, introduced originally as a device for getting research initiated, may result in empirical explanation.

Let me turn to another possible outcome of research models. Let us assume that the model is a statistical, mathematical, or mechanical one. If research shows that behavior corresponds to the statistical, mathematical, or mechanical laws, the investigator may now begin to think of and use the model as an explanatory system. The statistical (or mathematical or mechanical) laws may be used to deduce additional behavioral laws. Thus, if certain basic relationships between stimuli and responses in the area of learning were found to parallel those laws of inputs and outputs of electronic computers, then other relationships may be predicted for behavior by consideration of additional relationships which hold for computing systems. This procedure has not been very successful as yet in practice but a number of scientists in several different disciplines have an avowed faith that some such mechanical-electrical system as this will prove useful as an explanatory system for behavior. However, the main point I wish to make at this time is that this outcome of a research model is exactly the same as the initial reason for introducing models as given by other investigators. When a model is introduced, not merely to suggest research, but introduced with the avowed intent of using it as an explanatory system, we have the second general use of the word "model" which I call the explanatory model.

*Explanatory models.* In the physical sciences, there have been instances in which a mathematical system, originally developed without any reference to the real world, was taken over by a scientist and related to events in the world. That is, certain terms in the mathematical system were related to or identified with events in the world. If, then, it is shown that certain relationships among the mathematical terms also hold for the world events with which

those terms have been identified, the mathematical system becomes the explanatory system for the empirical events. And of course, manipulation of the symbols according to the rules of the system may lead to predictions of new relationships among events in the empirical world.

As indicated earlier, psychologists have often tried to emulate the behavior of physical scientists. We may expect, therefore, that psychologists might try to explain behavioral phenomena by using a mathematical system in toto, and indeed this has happened. For example, Lewin attempted to apply the system of topological geometry to behavior in a manner comparable to that indicated above for physical sciences (19). While this attempt has been judged to be mostly unsuccessful (9) for a number of reasons, I note it only to show that this kind of explanatory attempt has been sought after by psychologists. And, while the model was indeed unsuccessful as an explanatory one, it may have had some usefulness as a research model. I strongly suspect that the empirical content of an area must be fairly fully developed before the use of intact mathematical systems will prove very useful, if indeed they prove to be of any considerable use in psychology.

Nevertheless, mathematical symbols and the manipulative rules of algebra are being used in exploratory attempts, but on a very modest scale. Of course, the statement of behavioral laws in mathematical terms is the most precise descriptive technique we have for such laws. But, these mathematical model attempts are going beyond sheer description. First, a behavioral law, obtained in a simple experimental situation, is described by a mathematical equation. Then terms in the equation are identified either with known stimulus variables or with postulated processes, or both. Then, behavioral changes are investigated in other situations to see if these changes are predictable by the formulation. Now, actually, as it may be apparent, these explanatory attempts may be very close to the empirical-postulational and pure postulational explanations as discussed earlier in the chapter. I mention them here under the heading of models merely to emphasize the quantitative aspects under which they are initiated and which results in their frequently being called mathematical models. A more complete discussion of the role that mathematics

plays in such formulations can be found in several sources (e.g., 5, 25).

I have said that research models may be mechanical (or electrical, or other) models. So, also, may such models be introduced as explanatory models. That is, they are introduced not because the scientist wants to use them as analogies from which to develop research problems, but because he believes they will explain the behavior involved. This may take two distinguishable turns. Suppose an electronic computer is used as an explanatory system for learning and retention phenomena. As discussed earlier, if the laws for behavior and for the computer are commensurate, then the computer laws may be thought of as explaining the behavioral laws in the same sense that a mathematical system is said or might be said to explain behavior. Under such an orientation, one would not necessarily look (by research) for the neurophysiological counterparts of the computer; the laws for the computer and for the organism are postulated to be isomorphic and no inquiry is made as to how this comes about. In the same sense a mathematical model is postulated to be isomorphic with an area of behavior but it is meaningless to inquire as to how the mathematical system got that way.

On the other hand, the scientist may use a mechanical model as an explanatory model and then set about to find neurophysiological counterparts to the elements of the mechanical system. Actually, as we have discussed earlier this is a form of empirical explanation by identification, and occurs quite frequently in miniature form where psychology and neurophysiology converge, most notable in the areas of sensory processes and brain functions. For example, Kohler and Wallach (18) postulated a model for the visual cortex. This model was one of a particular kind of electrical field well understood by physicists. What these investigators said, in effect, was that if the cortex *was* such an electrical field then certain behavioral phenomena were understandable. That this model was not being used merely as a research model or as a formal model (as would be the case with a mathematical system) is shown by the fact that much effort of these investigators has been directed toward showing that such electrical fields do exist in the cortex. In short, it is thought of as a reductive explanatory model for behavioral phenomena. The model

was originally postulated but the research is directed toward a direct test of the postulated mechanisms.

I think I have covered the major usages of the term model in psychological literature. I have also tried to point out that the behavior of the scientist using models is not essentially different from the explanatory behavior as discussed earlier in the chapter. The scientist using research models is doing very much the same as is a scientist seeking empirical explanation, the difference being largely one of magnitude of the attempt. When a scientist uses an explanatory model he simply postulates that the intact model as already constituted has a counterpart in behavior. When, on the other hand, a scientist uses a postulational approach as discussed earlier he builds up the necessary concepts and relationships among them to serve as the organizing terms for the behavioral phenomena under consideration.

### CONCLUDING CONSIDERATIONS

We have examined a number of kinds of explanatory activities of scientists; we have seen the variety of concepts which may enter into explanatory attempts. We saw also the wide differences in opinion about when theory should enter into scientific activities and equally extreme opinions as to the nature of concepts which should be used in explanatory attempts. It think it is reasonably sure that whether some psychologists like it or not we will continue to have explanatory concepts introduced and used at all levels of abstraction; I think we may also anticipate explanatory attempts which will cut across areas in psychology as we know them today and across scientific disciplines too. So, what can we filter out of the previous discussion which will allow us to have a flexible attitude toward this somewhat chaotic situation, an attitude that is grounded in analytical thinking, receptive to new explanatory approaches, but critical of undisciplined thinking which leads to a proliferation of concepts held together only by the private intuitions of the writer who introduced them? How should one, in the initial stages of a research career, direct his explanatory efforts?

In discussing the methods of research we could be fairly firm (perhaps overly so) on matters of "good" and "bad" design. Can

we set down some rules for "good" and "bad" explanatory efforts which in no way hamper our conceptual imagination but will provide ground rules of some kind? It seems to me that the areas of psychology differ so in their empirical development that such working rules would be anachronistic in some areas and anticipatory in others. In the area of sensory processes almost any piece of research that is well done has relevance for explanatory attempts already put forward. In the area of clinical behavior we are so immersed in trying to establish reliable phenomena that explanatory efforts pretending any scope would be difficult to assess. In the area of learning, which lies somewhere between clinical and sensory processes in terms of empirical development, evaluation of explanatory efforts that lay claim to some scope have only recently been given systematic and comprehensive attention. An outline has been offered for evaluating such theories (9, p. xiii-xiv); some of the points in the outline have been discussed here but the entire outline deserves study by those who might be interested in the rather formidable undertaking of assessing systematically an explanatory system of some scope. Nevertheless, even when faced with my own argument against it, I have the temerity to make some comments and suggestions about explanatory procedures. Some of these, I feel sure, no one can disagree with; others might be decidedly controversial.

1. When reporting research I would insist that we have an obligation to place the research in some sort of context reflecting previous work. This context may take either of two forms. It may be a strictly empirical context in which the investigator makes an evaluation of just where the study fits, e.g., what gap is being filled, what empirical contradiction is trying to be resolved, and so on. This setting in the empirical context I judge to be obligatory. I have very little patience with research which is reported without reference to any other findings or other phenomena at the empirical level.

The second context in which an experiment may be presented (and this should be in addition to the empirical context) is an explanatory context. Many experiments are done for the purpose of testing hypotheses derived from some explanatory attempt. There are certain dangers involved in these experiments.

(a) The test may not lie within the boundaries specified by the explanatory system. A system developed to account for certain

operant conditioning phenomena only cannot be tested by using verbal learning materials. A positive result in such a case might suggest that the explanatory system could be expanded to accommodate a broader area of behavior, but a negative result (failure to support the hypothesis) must be interpreted with full respect for the boundary conditions or scope of the explanatory system.

(b) The "crucial" experiment is rare if not fictitious. Even the most unambiguously formulated explanatory system rarely leads to a prediction which a single negative experiment will result in overthrowing. A single experiment might require modification of a system but only a series of negative tests will require abandonment and there is an aphorism in science which essentially says that a theory is not overthrown by negative facts but only by a better theory. I think it is unfruitful to look for the will-o'-the-wisp crucial experiment; above all we should not expect that because, in our single test of someone's theory, we get findings contrary to the theory, the theoretician will say: "My apologies, sir, for advancing this theory; I'm obviously dead wrong."

(c) In general, as a safety device, we should try to design experiments so that the results will contribute substantially to the empirical base of our science even when our main intent is to test an explanatory idea. I recommend this because it so frequently happens that in spite of all precautions it will be discovered that an experiment may not be as relevant to an explanatory idea as was originally believed. If an experiment is designed to systematically determine the effects of a given variable or variables, but at the same time is thought to be a test of an explanatory idea, the data become a contribution to the science irrespective of how the theoretical aspects are judged. Except as an exploratory device in a new area, I think the two-condition experiment should be used very sparingly. So frequently when such experiments are designed to test theories they don't; they don't prove particularly evaluative of the theory and the data from them do little by way of sharpening theories. Explanatory systems must give laws as to how variables are related, not just state that they are related. The two-condition experiment can at best only tell us the latter.

2. I have indicated a number of ways by which explanation occurs in psychology. In all of these the scientist is looking for

generalizations; he is trying to bring phenomena and their relationships together in some meaningful fashion. But, just how does one go about this? Suppose that you have a set of reliable data in a given area and you want to bring this into some sort of explanatory order. How does one do this? I am in no position to tell *how* to theorize as far as the development of imagination, broad empirical perspicuity, and so on, are involved. However, in looking at the development of various explanatory attempts in psychology, I may suggest some alternative approaches with priorities, and some ideas on what to do irrespective of the nature of the specific explanatory attempt one chooses.

(a) I probably need not say, but shall anyhow, that we should carefully delineate the phenomena or relationships with which we wish to deal in our explanatory attempt. The phenomena should be operationally defined and their relationships with stimulus variables stated insofar as these are known. If the research has been systematic and precise enough, these relationships may be quantitatively stated.

(b) As a next step I think the search for empirical explanation has very high priority. This is most likely to be fruitful if one is dealing with data from a relatively new area of research. One asks whether or not the phenomena at hand can be manifestations of already established phenomena. This requires careful study of the operations but sometimes simple operational identification as an explanatory device is so obvious that it might be overlooked. It is my personal belief that no greater service can be rendered our science than by persistent attempts at empirical explanation. It may take additional research to establish the commonality of operations but this is true of any type of explanatory attempts. Now of course, if one accomplishes empirical explanation one may wish to pursue the matter further and offer "higher" forms of explanation (e.g., postulational) to encompass the "old" and "new" phenomena if an adequate system is not available for the old phenomena. This is a matter to which I will turn in a moment. What I would caution is that we don't jump to these higher forms of explanation without first considering carefully the possibility of empirical explanation. There are several illustrations in the literature of our science where a studious inspection of the operations defining a phenomenon would

have made empirical explanation very probable whereas we were given postulated processes instead. At least a consideration of empirical explanation may indicate additional research to define the boundaries of the phenomena which may then be brought under a postulated or postulated-empirical system. But this gets us back to the matter of theoretical readiness and I prefer not to let my own biases enter too strongly on this point since the merits of the alternatives are not in the least clear-cut and the level of empirical development in our science is very uneven in different areas. So, my main plea is to give empirical explanation a fighting chance.

(c) But if now, empirical explanation gets one nowhere, one may wish to introduce explanation of a postulated-empirical or postulational nature, whether in model form or not. Are there any cautions to guide us? I think there are. We certainly ought to first explore the possibility that an already existent system will handle the phenomena. Thus, even though we may not achieve empirical explanation, we may discover that higher explanatory systems will accommodate our findings. Our urge toward individuality should not blind us to a certain social responsibility for avoiding multiplication of concepts when this is not necessary. And yet, if one's conceptions of the processes underlying the phenomena are somewhat different from those implied by available concepts, introducing new concepts may be less confusing than using the old ones. This is a delicate problem and I know of no easy resolution. I think at the minimum one should state the similarities and differences in the characteristics of newly postulated processes and those which are already available. It may then become possible to resolve the differences satisfactorily so that only a single set of concepts will be needed.

Initially one may choose to present only a rough outline of his explanatory thinking with the intent of bringing precision to it later. In constructing any explanatory system alternative formulations are possible. Depending on your verve for living dangerously, you may or may not wish to do more research to choose among alternatives before stating the system precisely. It is clear that once one does offer seriously a postulational or empirical-postulational system it must not be ambiguous.

(d) It is a well-worn tenet that a theory should predict new facts; that is, facts over and above those used to induce the concepts in



the system. I suppose we should stick pretty close to this tenet, but I am a little resistant to it. A theory with postulated physiological processes might be untestable at the moment because of the state of technological development. That is, the instruments or other techniques needed to test the theory may not be available at the present time but five years from now they may have been developed. Even at the psychological concept level, one need not assume that a theorist is so infinitely wise that he alone is the one to determine whether or not his system allows for independent tests. Others might see how tests could be made. Nevertheless, I suppose we must continue to view not only the fact of whether or not a theory is testable but also the ease with which tests are generated, as prime criteria of theory evaluation.

3. It not infrequently comes to my attention that graduate students often try to establish their personal philosophies of explanation in psychology by asserting they are "for" or "against" theory. I wish the issues could be so simply resolved, but I think it is clear that they cannot. I think it is perfectly reasonable to expect to find certain explanatory methods, e.g., pure postulational, which are incompatible with one's mode of thinking. But, to say that one is against theory is not consonant with being a scientist. For although we couldn't arrive at any acceptable specific use of the word "theory" it nevertheless always implies an attempt to bring order to the world of empirical facts by abstracting out the commonalities underlying the facts. All this means is that we are searching for generalizations and this *is* science. When one asserts he is against theory it usually means he is against a particular way of approaching the search for generalizations, and that is the most it can mean if one is to remain a scientist.

## REFERENCES

1. ARCHER, E. J. Identification of visual patterns as a function of information load. *J. exp. Psychol.*, 1954, 48, 313-317.
2. BARTLETT, F. C. Programme for experiments on thinking. *Quart J. exp. Psychol.*, 1950, 2, 145-152.
3. BERG, I. A. Response bias and personality: The deviation hypothesis. *J. Psychol.*, 1955, 40, 61-72.

4. BROGDEN, W. J. Sensory pre-conditioning. *J. exp. Psychol.*, 1939, 25, 323-332.
5. COOMBS, C. H., RAIFFA, H., & THRALL, R. M. Some views on mathematical models and measurement theory. *Psychol. Rev.*, 1954, 61, 132-144.
6. DENNY, M. R. The role of secondary reinforcement in a partial reinforcement learning situation. *J. exp. Psychol.*, 1946, 36, 373-389.
7. DOLLARD, J., & MILLER, N. E. *Personality and psychotherapy*. New York: McGraw-Hill, 1950.
8. DUNCAN, C. P. On the similarity between reactive inhibition and neural satiation. *Amer. J. Psychol.*, 1956, 69, 227-235.
9. ESTES, W. K., KOCH, S., MACCORQUODALE, K., MEEHL, P. E., MUELLER, C. G. Jr., SCHOENFELD, W. N., & VERPLANCK, W. S. *Modern learning theory*. New York: Appleton-Century-Crofts, 1954.
10. FELDMAN, S. M., & UNDERWOOD, B. J. Stimulus recall following paired-associate learning. *J. exp. Psychol.*, 1956, 52, in press.
11. GLANZER, M. Stimulus satiation: An explanation of spontaneous alternation and related phenomena. *Psychol. Rev.*, 1953, 60, 257-268.
12. HEBB, D. O. *The organization of behavior*. New York: Wiley, 1949.
13. HELSON, H. Adaptation-level as frame of reference for prediction of psychophysical data. *Amer. J. Psychol.*, 1947, 60, 1-29.
14. HOVLAND, C. I. A "communication analysis" of concept learning. *Psychol. Rev.*, 1952, 59, 461-472.
15. HULL, C. L., HOVLAND, C. I., ROSS, R. T., HALL, M., PERKINS, D. T., & FITCH, F. B. *Mathematico-deductive theory of rote learning*. New Haven: Yale Univ. Press, 1940.
16. JENKINS, W. O., & SHEFFIELD, F. D. Rehearsal and guessing habits as sources of the "spread of effect." *J. exp. Psychol.*, 1946, 36, 316-330.
17. KENDLER, H. H. Experimental analysis of problem-solving behavior. Tech. Rep. No. 1, ONR, NR 150-064.
18. KOHLER, W., & WALLACH, H. Figural after-effects. *Proc. Amer. phil. Soc.*, 1944, 88, 269-357.
19. LEWIN, K. *Principles of topological psychology*. New York: McGraw-Hill, 1936.
20. MAZE, J. R. Do intervening variables intervene? *Psychol. Rev.*, 1954, 61, 226-234.
21. RIBBACK, A., & UNDERWOOD, B. J. An empirical explanation of the skewness of the bowed serial position curve. *J. exp. Psychol.*, 1950, 40, 329-335.

the system. I suppose we should stick pretty close to this tenet, but I am a little resistant to it. A theory with postulated physiological processes might be untestable at the moment because of the state of technological development. That is, the instruments or other techniques needed to test the theory may not be available at the present time but five years from now they may have been developed. Even at the psychological concept level, one need not assume that a theorist is so infinitely wise that he alone is the one to determine whether or not his system allows for independent tests. Others might see how tests could be made. Nevertheless, I suppose we must continue to view not only the fact of whether or not a theory is testable but also the ease with which tests are generated, as prime criteria of theory evaluation.

3. It not infrequently comes to my attention that graduate students often try to establish their personal philosophies of explanation in psychology by asserting they are "for" or "against" theory. I wish the issues could be so simply resolved, but I think it is clear that they cannot. I think it is perfectly reasonable to expect to find certain explanatory methods, e.g., pure postulational, which are incompatible with one's mode of thinking. But, to say that one is against theory is not consonant with being a scientist. For although we couldn't arrive at any acceptable specific use of the word "theory" it nevertheless always implies an attempt to bring order to the world of empirical facts by abstracting out the commonalities underlying the facts. All this means is that we are searching for generalizations and this *is* science. When one asserts he is against theory it usually means he is against a particular way of approaching the search for generalizations, and that is the most it can mean if one is to remain a scientist.

## REFERENCES

1. ARCHER, E. J. Identification of visual patterns as a function of information load. *J. exp. Psychol.*, 1954, 48, 313-317.
2. BARTLETT, F. C. Programme for experiments on thinking. *Quart J. exp. Psychol.*, 1950, 2, 145-152.
3. BERG, I. A. Response bias and personality: The deviation hypothesis. *J. Psychol.*, 1955, 40, 61-72.

## Potpourri

This is the final chapter. In the preceding discussions I have tried to exercise some restraint against continually inserting my biases into the exposition. I had hoped that I could mirror with reasonable accuracy some of the scientific activities of psychologists and not arouse too much intellectual caterwauling. Yet, since this is a personal document, undoubtedly distortions have been produced by my own convictions and by my own blind spots. In the present chapter I am making no pretense of reflecting anything but some of my own faiths, convictions, and prejudices. The points which I am going to discuss are not closely related; they are a miscellaneous group of issues which I choose to raise to the status of problems. Some of these have been touched upon at various points but I did not discuss them as fully as I wished because they were tangential to the matter of the moment.

### ANALYTICAL VERSUS NONANALYTICAL RESEARCH

I have said a number of times that the history of science is a history of relentless analysis. We aim to break down gross phenomena into subphenomena; we want to break complex stimulus conditions into their unitary parts. To those who insist that a whole is more than a sum of its parts we can only point out that justification for the insistence can eventuate solely from analytical research. Scientific advance depends upon analysis and inevitably follows the initial identification of gross phenomena with which the science must deal. In the material on experimental design I was harsh on nonanalytical research. I was, of course, speaking from the point of view which is looking for an understanding of behavior through knowledge of its basic relationships. What I wish to do now is pose this analytical research against nonanalytical research but in cases

22. SHIPLEY, W. C. Indirect conditioning. *J. gen. Psychol.*, 1935, 12, 337-357.
23. SPENCE, K. W. The differential response in animals to stimuli varying within a single dimension. *Psychol. Rev.*, 1937, 44, 430-444.
24. SPENCE, K. W. The role of secondary reinforcement in delayed reward learning. *Psychol. Rev.*, 1947, 54, 1-8.
25. SPENCE, K. W. Mathematical theories of learning. *J. gen. Psychol.*, 1953, 49, 283-291.
26. UNDERWOOD, B. J. An orientation for research on thinking. *Psychol. Rev.*, 1952, 59, 209-220.
27. WHITFIELD, J. W. An experiment in problem solving. *Quart. J. exp. Psychol.*, 1951, 3, 184-197.
28. WICKENS, D. D., & BRIGGS, G. E. Mediated stimulus generalization as a factor in sensory pre-conditioning. *J. exp. Psychol.*, 1951, 42, 197-200.
29. ZELLER, A. F. An experimental analogue of repression: III. The effect of induced failure and success on memory measured by recall. *J. exp. Psychol.*, 1951, 42, 32-38.

difference in performance between the two groups. Some of the more obvious ones are *knowledge of results*, *level of aspiration*, *competition*, *social interaction*, and even *chair arrangement*. Any one of these factors might have caused the difference between the experimental and control groups; or it might take two factors working together. It is the student's job, after evaluating this complex situation, to design a series of experiments in which the influence of each identified factor can be determined in isolation and the influence of combinations of these factors.

There are three points that I wish to make about an experiment in which a complex of factors constitutes the independent variable.

1. If the investigator is interested in demonstrating the influence of only one of these factors, then obviously this design represents an experimental error. We have discussed this matter at length in previous chapters.

2. Such nonanalytical research may have scientific utility in certain instances. In a research situation there are potentially many, many factors which might influence the behavior being measured. (I have made this statement so often in one form or another in the previous discussions that I feel confident it produces a certain amount of nausea!) To analyze carefully a given behavioral phenomenon, many variables must be independently manipulated. In the beginning of such analyses the investigator usually determines the influence of variables which he suspects are strongly related to the behavior and this is usually accomplished by isolating these factors in the experimental design. After a series of researches, however, he may have reason to believe that there are other factors which he hasn't specifically identified which are contributing to the behavior he is measuring. The basis for such a belief may come from several sources, such as discrepancies between his findings and other findings, certain theoretical expectations, suggestions of poorly understood interactions, and so on. Having no strong predilections as to which of several possible factors are involved, he may in a single experiment allow several of these factors to be operative simultaneously as in the illustration given above. Or, he may simply have a compulsiveness about ascertaining for sure that certain factors which he thinks are irrelevant are indeed irrelevant. In such a frame of mind the investigator might allow several of these factors to again

where the latter may have a very useful purpose and this purpose is so recognized by the investigator. Let me illustrate what I mean by nonanalytical research and then discuss this matter of utilization of such research results.

In the undergraduate course in experimental psychology which I teach I try to do at least one experiment which has features indicated by the following conditions. A simple task is used, such as cancellation, digit-symbol substitution, or reversed-alphabet printing. Two groups are matched on performance on the task chosen. Then the experimental group, working in isolation from the control group, is given a series of trials on this task. During these trials the subjects are seated in chairs placed in a circle so that each subject can see all other subjects. After each trial each subject counts the number of correct responses he made during the trial, e.g., the number of letters printed during a one-minute interval. When all have determined this value, the experimenter starts around the circle asking each subject to indicate clearly to all others how many he got correct and then how many he is going to "try for" on the next trial (level of aspiration). Then, another trial is given and after the number of correct responses is determined by each subject each in turn is asked to make known to all other subjects how many he said he was going to try for, how many he actually attained, and how many he is going to try for on the next trial. This procedure continues for several trials.

The control group, on the other hand, is simply given the equivalent number of trials in a formal situation. The subjects in this group are not allowed to count the number correct on each trial, are not allowed social interaction between trials, are seated in a formal classroom fashion, and so on. Comparison of performance on the series of trials usually shows the experimental group is superior; even these relatively sophisticated subjects usually respond to the conditions set up for the experimental group. But what do we have? We have a horribly nonanalytical experiment. Certainly the difference in performance can be attributed to the difference in treatments of the two groups. But, even at our rather retarded stage of knowledge in the area of human motivation we can identify a number of factors (operating in the experimental conditions and not in the control) any one of which could conceivably produce the

3. The outcome of nonanalytical research may have utilitarian value over and above any considerations of its scientific value. If a complex of factors is known to increase motivation this might have some use, say, in the schoolroom. Information concerning how much technical knowledge is remembered by a radar repairman one year after completing school may be very valuable data without having any information about which of many possible factors was responsible for the forgetting. I shall not dwell on this matter; I point it out here only to suggest that the criteria of good science do not necessarily overlap with those of useful data.

### SOME MORE BIASES ON DATA ANALYSIS

I hope I can say what I wish to say here without sounding pontifical, for my own research practices are still too far separated from the dogma I wish to present to justify my assuming the attitude of a prosecutor, even if the dogma is "correct." But, without further apologies, I would like to state simply that a major weakness of current research in psychology is that many data which have been collected are inadequately analyzed and this has resulted in gross inefficiency in our total research effort.

An experiment is usually set up to test one or two hypotheses or to determine the influence of one or two variables on behavior. Far too often we analyze our data just enough to test these one or two hypotheses or to state the relationship between the variables and behavior. I insist that most researches will accomplish much more than this if we just give them a chance to do so by studying our data in many different ways over and above those dictated by the problem or hypothesis which the experiment was designed to test. I will expand these statements by a series of points.

1. Progress in advanced forms of statistical analysis, particularly analysis of variance, has had two major impacts on data analysis. First, it has allowed a means of testing the significance of behavioral interaction resulting from the combined influence of two or more variables. Secondly, it has aided in diminishing the reasons I can cite for advancing the criticisms of nonanalytical experiments in the previous section. The influence of many variables, including those not necessarily judged to be particularly relevant, can be evaluated



be operative in a single experiment. If no difference in behavior occurs as a consequence of the manipulation of this complex stimulus situation the investigator has in a single experiment eliminated three or four (or as many as he can identify) variables as being relevant for the behavior being studied. There is some danger, of course, that say, two of the variables might have opposite effects and thus cancel, but usually a careful internal analysis of the data can detect such a possibility. If, on the other hand, he does find a difference in behavior, he knows the critical variable is among the several involved, or he knows that there is a combination which influences behavior, and he can proceed with analytical research to isolate the factor or factors involved.

The above procedures are not frequently used intentionally in psychological research; perhaps not as frequently as they should be. I think such research might be quite efficient under circumstances either where the investigator judges the variables not to be relevant but wants this "on the books" or where he has reason to believe that there might be an important variable among several possible ones. And I might add that as I shall point out more specifically later, there are many instances in our science where an investigator prejudged a variable as not relevant only to find that it was highly relevant.

I said above that it doesn't seem to me that such procedures are used very frequently in psychological research. The investigator usually doesn't indicate whether or not he realizes he has several possible unitary factors involved in the stimulus complex. Clearly, if we are using this method as a means of exploring several variables we should indicate this in order to avoid being criticized for lumping all these factors together. The elimination of variables as relevant factors for behavior is, of course, a very worthwhile scientific enterprise since it is an integral part of analytical progress. While there seems to be no *a priori* reason why we should ape the older sciences, it is my understanding that this shotgun approach for demonstrating the irrelevancy of factors is used in these sciences and we might well study further the implications of such emulation. I am going to return to this matter very briefly in other contexts in this chapter.

cifically done. Obviously, many subject variables fall into this category, such as age, sex, educational status, and so on. But there are many other environmental and task variables, which we judge to be unimportant, when we ought to go about demonstrating this. These include influence of time of day when the subject is run, the influence of different experimenters, and many, many others which could be specified after analyzing a particular research situation. In short, some day we have to determine the effects of many variables, and whether we think they are relevant ahead of time or not, the determinations will have to be made. Many of these assessments can be made as subsidiary endeavors to the major purpose of the experiment, thus avoiding the necessity of having to design experiments for the express purpose of investigating them.

I appreciate an attitude reflected by a statement: "But I am not interested in those variables," such a statement being given in response to the above. The lack of interest may come from the fact that no theoretical import is attached to the variables. I can only say that the history of science shows that some variables which were presumed to have no theoretical import turned out to have one; variables which were not even thought to be empirically relevant turn out to be so and then gain theoretical relevance; variables which have no relevance to one's particular theory may have it for another's theory. So, I can only base my plea on the grounds that each scientist has a responsibility to the science in which he works and the analyses that I have suggested may help to discharge some of this responsibility.

As I have said, I am far from satisfied with my own efforts in this direction, but I have done enough slicing of data to know that the dividends can sometimes be astonishing. I have made gross errors in judging the probable relevance of variables. These errors were discovered by making what I thought would be a routine subsidiary analysis with full expectation that the analysis would show that the variable was of little consequence, but as it turned out it was a highly relevant factor. The discovery of interaction between stage of practice and degree of intralist similarity and its influence on retention came about in this fashion (2). Perhaps I should have guessed this interaction would be present, and certainly after it was discovered it seemed obvious to me, but the fact is I didn't think it was im-

by including them as separate terms in the analysis. Thus, such analyses may achieve the same end as I have suggested would be achieved by throwing several variables into a stimulus complex. As a matter of fact, it is quite clear that considered use of analysis of variance can achieve this much more efficiently. But, while it is trite to say so, one gets out of an analysis of variance only an evaluation of those variables which are put into it and what one puts into it depends not only on statistical and design acumen but also psychological acumen concerning the range of factors which might influence the behavior. Furthermore, the subsidiary factors often inserted in analysis of variance are inserted not because of an interest in their influence or lack of it but because a purer estimate of the error term for testing the major effects results from this insertion. The same factors so often become standard from experiment to experiment that the potentially powerful tool is not given a chance to evaluate the influence of new variables that have been prejudged to be subsidiary.

Regardless of how comprehensive an analysis of variance may be, I suspect that it is always possible to look at the findings from a different angle; it is always possible to slice, fractionate, or combine in new ways to obtain more information. And this is all I am urging; I think we should ask ourselves many, many questions about any set of data and see if these questions can be answered with the data at hand, thus avoiding having to do a new experiment. I am aware of the aversion that statisticians have for testing successive *ad-hoc* hypotheses from a set of data; but this fear is groundless if the investigator evaluates tests of these hypotheses with a judicious consideration for the statistical issues.

2. I mentioned that certain standard analyses of variance patterns do provide tests of the influence of variables which the investigator would probably guess beforehand were probably not of much moment for the behavior being studied. It is simply the responsibility of the investigator to put these into the design. Without additional reference to analysis of variance, I would like to discuss further this matter of determining the effect of many variables. Almost any research offers the possibility of answering questions concerning the influence of certain variables for which the research was not spe-

## THE ORTHOGONAL DESIGN .

Let us face the fact that a great deal of research is done by plodding pedants, among whom I am not at all unhappy to place myself. Many of us simply do not have that elusive skill for uncharacterized capacity which somehow sweeps away the unessentials and leaves stark before us the fundamental empirical principles of a science or an area within a science. Nor do we have the ready skill to see how these principles can be related by abstractions which seem to spring from others with irritating ease. Our empirical generalizations stem from the tedious work of the laboratory; our useful abstractions which relate them come, if they come at all, only by tortuous trial and error. Our satisfactions come from the beauty of a systematic relationship, from observing frequently how lawful the behavior of the living organism is, and from devising designs which will lay bare the essential components of a complex situation. As we work we may develop some pretty strong ideas as to what procedures will move us along most rapidly toward other intriguing problems which we see but to which we feel we should not jump until the problems at hand have in some sense been solved. We develop a certain antipathy toward (but at the same time an envy of) the peripatetic investigator who jumps nimbly from area to area, who never goes deeply because he is so continually buoyed by the elixir of discovery. We sometimes wish this investigator would settle down and help us by developing systematically some specific area. Our better judgment would not ask for this, however, not only because we would insist on freedom of inquiry but because we realize that this happy vagabond may just possibly discover an important key to understanding behavior.

So, having fallen into a role of the plodding pedant (undoubtedly resulting from some multiplicative function of philosophical conviction  $\times$  skills  $\times$  temperament) we, as I said earlier, develop some firm convictions about the nature of research designs which will facilitate our analyses. One of these convictions which I hold is that, at our present stage of knowledge of experimental design and statistical analysis, the orthogonal design gives us the most information in the most efficient sort of way. There are several points I wish to make about this type of design, but I will work into them gradually.

portant and we probably would still remain oblivious to it if we had not established a habit of making many subsidiary analyses of data in our laboratory. I do not mean to imply that this finding had any earth-shaking consequences, but in my own very restricted area of research it allowed a lot of puzzling findings to fall in line.

3. Sticking too close to analyses dictated by the problem for which the experiment was designed has other unsatisfactory consequences. There are a number of published reports in the literature which take the following pattern. A problem is stated, the results analyzed around this problem, and a theoretical interpretation of the results is offered. Then, it is usually stated that subsequent research will have to demonstrate whether or not this theoretical interpretation is useful. But, if the investigator studied his procedures and data carefully he would find that he already has data from the same experiment which suggested the explanation which would test the theory. But not perceiving that this is possible, he will go ahead (or someone else will) and do a new experiment to test the hypothesis which could have been tested by analyzing the data in a different way from the already completed study. I think this is wasteful of research energy. Of course I realize that support for the theory is of little consequence since it comes from the data which suggested it, but I am concerned with data gotten from subanalyses which are not in line with the theory. For, if such analyses had been made the interpretation actually made would not have been given in the report.

Allied with this matter is another. As our thinking develops in a given area we may get explanatory ideas, or ideas for the importance of certain variables hitherto believed irrelevant, and we may have data available in the files to make at least a preliminary test. One of the cardinal sins in our laboratory is to throw away raw data once the major analysis and whatever subsidiary analyses we thought of at the time are completed. It may be three years, or ten years, or never that we arrive at a point in the development of an area where we ask questions that we judge important and which can be answered by subsidiary analysis of data already in our files. Many times these working hypotheses have turned out to be incorrect, but we have determined this without having to run a completely new experiment.

pretation of this complexity is that there is an interactive influence of variables on behavior, which indeed there is. But, I think we may be ultimately surprised that the interactions are less frequent than we lead others to believe. One need only leaf through a standard journal reporting research where orthogonal designs are frequently used to discover that many, many interaction effects are not significant statistically when *a priori* consideration of the variables involved would lead one to guess they would be. In short, it is barely possible that we are overstating the case for complexity when this implies interaction phenomena. However, the main point I wish to make is that since we now have the statistical tools for examining interaction, we should make the orthogonal design a standard practice (preferably a minimum of a  $3 \times 3$ ) and find out as rapidly as we can just what complexities we have to deal with. We have little to lose by these designs, and much, very much, to gain.

3. One of the obvious drawbacks to the orthogonal design is its extensiveness. So, even when using such designs there has been a noticeable tendency to assign only a few subjects to a cell. Statistically speaking these small numbers can be justified, but lawfully speaking, when the numbers get small the relationship between each variable and behavior becomes quite imprecise. And in many areas I think we should be getting fairly accurate estimates of our relationships. Furthermore, the smaller the number of subjects in a cell the less the opportunity to slice the data meaningfully in ways different from that originally intended by the design. I am referring, of course, to the subanalyses of data as discussed earlier. If we can avoid it, therefore, let us not allow statistical justification to blot our major objective of understanding behavior which comes first through the experimental derivation of laws.

### STANDARDIZATION

The issues (if they be issues) suggested by the topic of standardization are not easy to delimit and specify. We have little by way of written points of view on the topic; most of them have come to my attention through oral communication. Melton (1), in 1936, evaluated the status of methodology in human learning and made positive statements about the benefits of standardization. Since this work

1. Suppose we wanted to find out the influence of a given variable on a given response. We may, at the very simplest level, want to know only whether or not it is a relevant variable. So, we might use a simple two-condition experiment, choosing some value of the variable for the experimental condition and zero value for the control. This allows operational definition of the phenomenon if there is one. Under usual circumstances we would probably get an answer to the question of whether or not the variable is relevant for the behavior being studied. But, negative results are not very convincing. We might have sampled the "wrong" value of the variable; its true relationship with behavior may be such that only small amounts or medium amounts will influence behavior.

Now suppose, in view of the above considerations, we sample two values along the dimension, and historically this usually means two extreme points. We are reducing rather greatly the probabilities that we will hit a "dead" spot, although in certain areas of behavior we might quite easily. But, even if we get positive results (if the results show the variable is relevant) we can say very little about the nature of the relationship. The relationship might be curvilinear, negatively accelerated, positively accelerated, linear, and so on, and we would not know it. Now, if we tap three points along the dimension, say at the extremes and in the middle, the amount of information added is tremendous. We can be fairly confident that if we get no difference (over and above the control) the variable is not relevant, for it would be highly unlikely that we hit three dead spots. Furthermore, if it is a relevant variable, we have a fairly accurate estimate of the nature of the relationship although no one would deny that adding more points increases our confidence in this matter. So then, my first point is that except for pilot studies, if we are seriously asking about the influence of a given variable, we should tap the stimulus variable at least at three places, ideally widely spaced.

2. The orthogonal design explores the influence of at least two variables simultaneously. If we use three values along each dimension the result is two sets of relationships (as discussed above) plus the interaction effect (if any) of the two variables. The determination of interaction effects is important. I have a feeling that we have too long hidden ourselves behind the oft-given statement that behavior is so complex that we just can't make progress very fast. One inter-

and there is no way by which this instrument *per se* can vary from laboratory to laboratory. However, there are many potential variables for which it is difficult to attach values so that equivalence could be maintained from laboratory to laboratory. One of the most obvious sets of variables is subject differences, whether these be rats, monkeys, or men. Or take such a simple matter as differences in smells in animal experimental rooms; or the difference in volunteer subjects *versus* paid subjects; the difference in color of walls of the experimental rooms, and so on. These factors could possibly be assigned values so that replication is possible but in practice it is very difficult to do. Presumably, using the same strain of rats might result in standardization but has anyone shown that rats used at Yale from Strain X behave in the same way as the rats from Strain X at Oxford? This situation has led in some cases to research that is called methodological research. All this means is that we will run experiments to find out if differences, say, in experimenters, or differences in educational level of college students and so on are relevant variables. This research is no different than any other kind of research; if the variables are shown to be irrelevant for certain phenomena then one need not worry about the standardization of this variable when subsequently investigating these phenomena. If a variable is shown to be relevant then, by the idea of standardization, we should find means of holding this at the same value from laboratory to laboratory.

2. What is to be achieved by standardization? There are several objectives which one could state but they all end up resulting in some sort of unification of knowledge. Data obtained in one laboratory may be immediately unified or merged with data from another. A systematic body of knowledge is built up and we avoid the likelihood of isolated bits of research which may not fit anywhere. It might be said that standardization is more efficient than haphazard research, even though the designs representing both may be equally good as far as accepting conclusions from them is concerned. Standardization will lead to a body of internally consistent relationships in the sense that all were evolved from the same basic set of conditions and theory construction can proceed with full confidence in the reliability of the relationships. To be sure, the analysis and under-



is 20 years old and deals only with human learning, it would be quite unfair to use it as representative of contemporary thinking (of Melton and others) even though it may well be. So, I shall state a position which probably represents some scientists' point of view and then I will order the discussion around the implications of the statement.

Suppose you were going to do a study in an area in which considerable work had already been accomplished. A program of standardization would require that unless you had specific reasons for doing otherwise you should keep the values of all static variables equivalent to those values used in previous studies. By static variables I mean those factors which are known to influence the phenomenon under consideration or which conceivably could so influence but where their relationship to the phenomenon under consideration is of no concern for this research. Thus, if you were going to perform a study on tachistoscopic thresholds of verbal material as a function of meaningfulness of words, and if most other research had used the ascending method of limits, increasing exposure time rather than brightness, and so on, the principle of standardization would say that these static variables should be the same as used in other studies. The series of points to follow concerning this matter will suggest implications of such procedural conformity, both positive and negative, as I see them.

1. I have used the term "value" above, saying that the value of static variables should be kept the same as they had been in previous studies. Value thus implies a quantification of some sort. Such duplication of quantitative values when physical scales are used would seem to be relatively straightforward but it isn't always. Obviously, until standardization of calibration is achieved in such instances one cannot, even if desired, keep the static variables equivalent from one laboratory to another. I don't think anyone would deny that it is a sad state of affairs when one wants to achieve standardization, thinks he is, but actually isn't.

When we are dealing with variables whose characteristics cannot be reflected by a physical scale, but which must be reflected by a psychological scale, we are in certain respects better off than when a physical scale is used. Thus, we can use nonsense syllables of specified degrees of meaningfulness as scaled by a particular investigator

phenomena in detail. Research gets built up around certain behavioral phenomena. The research problem is to determine which variables do and which do not influence these phenomena and for those that do, to determine the form of relationship. Phenomena gain different levels of generality depending upon how diverse the situations are in which they will occur (even though different situations may result in different amounts of the phenomena). This whole process is facilitated by empirical explanation as discussed in the previous chapter. But, the fact that a phenomenon is a very general one does not remove our responsibility to determine whether variables other than the critical defining one do or do not influence the magnitude of the phenomena. How do we go about determining whether this great host of potential variables are or are not relevant in that they change the amount of a given phenomenon?

In the beginning of research on a phenomenon the investigator usually makes guesses as to which variables will influence behavior, these guesses being perhaps suggested by his research model or perhaps by some sort of articulate theory carried over from another area. Usually the investigator starts out by examining the influence of variables which he thinks *will* affect the phenomenon; he doesn't usually start by manipulating variables which he thinks *won't* have an influence. Models and theories usually don't predict irrelevant variables. The guesses concerning the relevance of variables are far from infallible; if they were infallible we would need the research only to determine the precise laws. But now, let us get down to the point of standardization. A scientist who favors standardization and one who doesn't would probably both agree that the basic aim of science at the empirical level is that of determining the generality of phenomena and the variables which determine the amount of the phenomena. Let us see, then, how this objective may be reached by research conducted under the aegis of standardization.

First, let us not take an extreme case of standardization; we do not need this to heighten the points of view. Let us take the case where an area has had some preliminary working over; it appears to be a fruitful area and so we want to explore systematically the effects of variables on the phenomena representing the area. We recognize a great many possible variables which could influence the phenomena; but we prejudge the situation and make guesses as to the

standing is achieved within a very limited context but it may later be broadened beyond this.

3. Now, why is there an issue involved? Who could object to such a program of standardization? Let me first dispose of one objection which is occasionally raised, namely, that the principle of standardization is contrary to the principle of investigative freedom. Such a criticism seems to me to result either from a misconception of what is meant by standardization or a misunderstanding of freedom of inquiry. There is nothing in a principle of standardization which limits one's area or which prevents him from exploring the effects of any variable one wants to. All the principle says is that we should have continuity of variables if we are not interested in the influence of those variables. I don't see how even the most rabid protector of research freedom would say that this abrogates or threatens this freedom. The principle says that here are factors in which you are not interested; wouldn't it be worthwhile to handle these in such a way so that different researches in different laboratories may become a common body of data? It seems to me that if anyone would object to this on the grounds of a threat to research freedom alone he is showing somewhat irresponsible behavior.

4. I think objections may be raised to the principle of standardization which are serious, legitimate, and which must be considered before one embraces the principle. The objection comes from a consideration of the objectives of science and how those objectives can be most efficiently achieved. To advance the argument systematically will take a little preparation.

With some fear of reprisals, I must say again that we have noted that at the empirical level the number of variables which might influence behavior is very great. Actually, there seems to be no way to avoid in the long run of our science the task of somehow determining which factors are relevant and which are not. It is complicated further by the fact that as we well know a variable may be relevant for one form of behavior and not for others. Then there are the interactions about which we have spoken. In the face of this gigantic task one might throw up his hands in despair; indeed, some have done so and taken up other pursuits. But, there are many others who have gone about their research, got hold of a restricted range of phenomena, and have gone about the analyses of these

he would do to get the same information and I very likely would do a lot less. In short, it seems to me that the goal of generality—the generality of phenomena and laws about them—can be most rapidly achieved by not following a dictum of standardization. I think this has happened and is happening in psychological research quite by accident and by the exigencies existing in various laboratories. I would not like to see this replaced by standardization. But, there are two additional comments which must be made.

The accretion of generality by nonstandardization cannot be accomplished if differences in research procedures cannot be specified. If differences cannot be specified then one is quite unable to pin down discrepancies. I think sometimes that those who speak most strongly for standardization really mean only this; if so, I cannot disagree.

The second point is that within our own laboratories a certain amount of standardization is evident. If a series of researches is done it is often done in just the way that a devotee of standardization would have it done. But I think this happens because of convenience and not because of philosophical convictions. And, as is evident, I am by no means convinced that this is what we should do even in our own series of researches.

### REPORTING RESEARCH

From the reading of many manuscripts reporting research, I have arrived (along with many others) at the conclusion that as a profession we are not very good at writing scientific prose. Too, I have seen Ph.D. dissertations go through several drafts, not because the research wasn't well done, but because the candidate just couldn't set down on paper what he did, what he found, and what he made of it. At one time I had hoped I might be able to bring together certain suggestions for writing research reports which would make this matter of writing scientific prose a less burdensome activity for thesis advisers and journal editors. But I find I cannot do it. When I put them down on paper it sounds as if I were telling a stutterer to stop stuttering without telling him how. All I could possibly have said would be such things as "write clearly," "make your points sharp," "define your terms," "have a definite progression," and so

variables which are likely to be most relevant. We then start about our business of isolating the effects of these variables. All others, including those which we judged to be irrelevant are fixed at a given level throughout the series of researches; they become the static variables of the situation. It is quite obvious that a tight, highly systematic body of data could be built up around this situation. Now, let us suppose that an investigator at another university gets interested in the area of research. The standardization dictum would say that he should duplicate the values of the static variables of the first investigator. If he did this, the two sets of data could be readily merged.

The generality of the phenomena studied within the highly restricted set of conditions as outlined is unknown. But this generality could be determined over a period of time by ticking off one variable after another by a series of researches; that is, one could now investigate the influence of variables which had previously been static. Subject characteristics may be varied, task characteristics and so on. But I would venture an opinion that this procedure is inefficient. If I take a position that we should not standardize in this sense, and I do take such a position, I advance the following sorts of arguments.

We prejudge many stimulus conditions to be irrelevant and we are frequently wrong. Supposing that I, as the second investigator in the above situation, did not adhere to the principle of standardization. So I do an experiment which repeats one of the researches of the previous investigator (at least in part) except that I deliberately use different values of, say six static variables, whether I judge them to be relevant or not. Now supposing I obtain the same results the previous investigator obtained. Barring the cancellation effects of variables having opposite influences, what have I achieved? It seems to me that as discussed earlier I have shown in a single experiment what would have taken the other investigator six experiments to show if he adhered to his philosophy. If, on the other hand, I do indeed find a difference between his results and mine, I have identified a pertinent variable which he may have judged, at least tentatively, to be irrelevant and I can now proceed with more analytical research to determine which variable or combination is responsible. Under no circumstances would I ever have to do more work than

the research I did. I realize the value of replication of research but I would like to know when I am replicating; many researches do not need replicating directly. I am told that there is a scientist at one of our large universities who has done many well-executed experiments on learning; none of it is published. I do not believe that a private science goes along with freedom of inquiry. If a person does a piece of research that is sound it seems to me that he has a responsibility to make this available to all, if for no other reason than efficient advancement of our science. Of course we all do research which we decide should not be published for one reason or another (or an editor decides for us) but this is a decision to be reached for each research project and should not become a way of life.

I am fairly sure that many scientists lose the fun of research once the data are in and analyzed and that this is responsible for failure to publish significant data. Writing research reports is a chore for many, and unless pressures are exerted the data remain in the file. As most in the profession know, promotions are in many institutions based on production of research publications. There may be many evils attached to such a policy but certainly one positive aspect is that it may prevent excellent data from remaining forever in files.

But let me be more positive about this matter of writing up completed research. It has been my experience that the analyses which are done before one starts to write the first draft of a report are seldom all of the analyses which are eventually done. When one is forced to put his thoughts in writing and to evaluate the implications of his findings precisely, it often happens that other analyses are suggested. Occasionally these additional analyses turn out to be quite important. Even this may not be the end, for when an editor reads the manuscript he may get ideas for further breakdowns or other analyses which will support, modify, or recast the findings. Thus, it seems to me that research reporting is simply a further step in scientific thinking and is not merely a chore. But one has to write before these later benefits will accrue. I am sure that the motivational factors involved in publishing or not publishing research are much more complicated than I have indicated. All I can say is that I think we have an obligation to publish sound research

on. I find such suggestions quite revolting. I have therefore concluded that there is little I can say in a positive way which I can condone and so am left moaning over our writing ills.

But, still talking at the level of platitudes, I am convinced that writing scientific prose is a skill which develops with practice and knowledge of results provided by someone who is judged to be able to write straightforward research reports. Our conviction in the worthwhileness of this platitude is strong enough so that both our undergraduate and graduate students are given practice in writing manuscripts (over and above theses) and these are meticulously marked for clarity of expression, organization, and so on. It would be nice to have a control condition for this treatment; if we did I suspect our faith in the value of the practice might be considerably shaken. Nevertheless, I suspect there are many worse ways for students and instructors to spend their time in a program designed to train scientists.

There is an issue, however, regarding the reporting of research, which I feel is worth a little space. Some of the issues on which there are varying shades of opinion which we have discussed in this book were resolved by appealing to a criterion of efficiency in our scientific pursuits. I don't know for sure that efficiency is a legitimate criterion to use in deciding issues even if it is the only differentiating factor between two positions. I have obviously used it in some cases to aid in arriving at a point of view. This does not, of course, give it a sanctified status even though our culture nearly forces such a status on us. Nevertheless, since science is a part of our culture, and since saving in manpower may result from efficiency in science as well as in an A & P store, I do not think we should disregard it as a criterion. Be this reasonable or not, I would like to mention briefly certain ideas about the scientists' obligation to report research.

If our universities grant us time, money, and freedom to do research as we choose, we have certain responsibilities in return. One of these obligations is that of making available to the scientific public the results of our investigations. I frankly get a little disgusted hearing by word of mouth that so and so did an experiment five years ago and found exactly the same as I found in an experiment just completed. The other investigator never got around to publishing the results of his research; if he had I never would have done

relatively small. I am not, of course, making any value judgments about such areas; they are, for all I can tell, getting at very important behavioral processes. I mention them as possible danger signals that research facts in some areas may be accumulating faster than they can be assimilated. Until approximately 10 years ago the *Psychological Bulletin* served an extraordinarily useful function of keeping the research psychologists up to date on the theoretical and empirical content of all areas in psychology. I have been told that in recent years editors of the *Bulletin* have had increasing difficulty in obtaining reviews in some of the older areas of research because of the massive task which faced the reviewer. A book which summarized the area of conditioning in 1940 has remained unrevised because, apparently, of the same problem.

It would seem to me, in view of these signals, that we must reward more highly the empirical integrator. This would be a man who, in certain respects, fits the conception of the humanist scholar. His research work will be done in the library; he sorts research findings into piles according to an integrative scheme; he notes which findings should *not* be handed down to successive generations because of flaws in the research methods; he summarizes the sound empirical findings; he notes contradictory findings and makes guesses as to the reason for the contradiction; he points out gaps in research; he may note the status of explanatory efforts. Thus, several hundred research reports may be reduced to a report that can be easily assimilated. I think we all know scientists who could do such work with skill and who may or may not themselves be active research workers. I furthermore suspect that, at least in certain areas, this work would be far more valuable to our science than the actual research done by the man. The data he saves from oblivion-by-neglect may be far more valuable than the data he would collect.

I guess what I am trying to say is that I do not want to leave the impression that advances in science are made only in the laboratory or in the theoretician's office. The empirical integrator is very much in demand.

. . . . .

I have been calculating that I have a probable allowance of 25 years of research energy remaining. I must waste no more of this



findings; whatever the motives are which lead to this I am for; those that prevent it, I am against.

### THE EMPIRICAL INTEGRATOR

My major point of departure in this book has been problems facing the worker actively engaged in planning, executing, and interpreting research. It is from his activities, when properly carried out, that we gain initial status as a scientific discipline. But, data, facts, and relationships cannot be allowed to lie undigested on the pages of our journals. One of the purposes of theory is to provide this assimilative function. But I do not believe that explanatory attempts are satisfactorily handling this function at the present time. If I correctly assess the current situation, it is that we have vast bodies of data even within areas within psychology which desperately need to be brought into some sort of integrative scheme. And the data are being spewed out at ever-increasing rates. What are the dangers of this?

In some of the older sciences, great quantities of research findings remain unincorporated within theories. In certain areas within these sciences there is grave danger that many of these data will be lost to successive generations of students. When such mounds of data remain unintegrated no graduate student can, without a nearly prohibitive amount of time and effort, be expected to master these facts. Consequently, the student tends to shunt himself toward more newly developed areas where the background study for his own research can be accomplished with the expenditure of a reasonable amount of time and effort. The upshot of this is that some of the older areas of research are no longer attracting research workers, not because the problems in those areas are unimportant, but because it is so difficult "to get ahold" of the area in a manner befitting a scholar.

If I read the signs correctly, certain areas of psychology may be approaching this point. I note what might be called "fads" in research areas develop and that we will have a spate of Ph.D. theses in these areas. Thus the work on perceptual thresholds and probability learning seems to fit this category. The amount of background reading necessary to do well-informed research in these areas is

# AUTHOR INDEX

If author's name does not appear on page in text where reference is cited, the reference number appears here in parenthesis after the page number

- Abelson, R. P. 45 (1), 48  
 Adams, D. K. 231, 232  
 Adamson, R. E. 163 (1), 171  
 Allport, G. W. 18 (2), 48  
 Alper, T. G. 95 (12), 126, 153 (2),  
 160 (2), 171  
 Ammons, H. 73 (1), 83  
 Applezweig, M. H. 102 (1), 126  
 Archer, E. J. 259 (1), 268  
 Arnoult, M. D. 138 (3), 171  
  
 Baker, K. E. 137 (17), 172  
 Barker, R. G. 132 (4), 171  
 Baron, M. R. 104 (3), 126  
 Bartlett, F. C. 258, 268  
 Beck, L. W. 178 (1), 193, 203, 223 (2),  
 233  
 Beck, S. J. 88 (2), 126  
 Belmont, L. 149 (5), 171  
 Benjamin, A. C. 5, 16  
 Berg, I. A. 246, 268  
 Bergmann, G. 52, 83, 177, 179 (2), 193,  
 231, 233  
 Bills, A. G. 53 (3), 83  
 Bilodeau, E. A. 104 (3), 126  
 Birch, H. G. 149 (5), 171  
 Bobbitt, J. M. 66, 83  
 Boring, E. G. 130, 172, 178, 193  
 Bousfield, W. A. 55, 83  
 Braithwaite, R. B. 179 (5), 193  
 Briggs, G. E. 243 (28), 270  
 Brogden, W. J. 29 (3), 48, 185, 193,  
 242 (4), 269  
 Brower, D. 133 (7), 172  
 Brown, J. S. 104 (3), 126  
 Bruner, J. S. 139 (8), 172  
 Busiek, R. D. 139 (8), 172  
  
 Campbell, D. T. 145, 147, 172  
 Campbell, N. R. 177, 178 (7), 193  
 Cattell, R. B. 32, 33, 49  
 Chandler, K. A. 108 (17), 109 (17),  
 110 (19), 126, 127  
 Chapin, F. S. 39 (5), 49, 98 (4), 126  
 Child, I. I. 132 (10), 172  
 Clark, W. H. 105 (5), 126  
 Cohen, M. R. 176, 178 (8), 193  
 Combs, A. W. 133 (11), 172  
 Cook, S. W. 36 (15), 49  
 Coombs, C. H. 19 (6), 49, 262 (5), 269  
 Cowen, E. L. 133 (11), 172  
 Cronbach, L. 118, 126  
 Crown, S. 133 (12), 172  
  
 Dallenbach, K. M. 191 (9), 193  
 Davis, R. C. 232, 233  
 Dembo, T. 132 (4), 171  
 Denny, M. R. 241 (6), 269  
 Deutsch, M. 36 (15), 49  
 Dollard, J. 24, 49, 248, 269  
 du Mas, F. M. 89 (7), 126  
 Duncan, C. P. 27 (8), 49, 246 (8), 269  
  
 Edwards, A. L. 153 (13), 160 (13), 172  
 Elkin, A. 30, 49  
 Estes, W. 190 (10), 193, 258 (9), 261  
 (9), 264 (9), 269  
 Eysenck, H. J. 33, 49, 118, 126  
  
 Farrell, B. A. 230, 233  
 Feigl, H. 3 (2), 16  
 Feldman, S. M. 237 (10), 269  
 Fitch, F. B. 238 (15), 253 (15), 269  
 Fleishman, E. A. 170 (14), 172  
 Fredericson, E. 143 (15), 172

allotted time behaving like a venerable seer, for that is what I fear I have been doing in this book.

## REFERENCES

1. MELTON, A. W. The methodology of experimental studies of human learning and retention. I. The functions of a methodology and the available criteria for evaluating different experimental methods. *Psychol. Bull.*, 1936, 33, 305-394.
2. UNDERWOOD, B. J. Intralist similarity in verbal learning and retention. *Psychol. Rev.*, 1954, 61, 160-166.

allotted time behaving like a venerable seer, for that is what I fear I have been doing in this book.

## REFERENCES

1. MELTON, A. W. The methodology of experimental studies of human learning and retention. I. The functions of a methodology and the available criteria for evaluating different experimental methods. *Psychol. Bull.*, 1936, 33, 305-394.
2. UNDERWOOD, B. J. Intralist similarity in verbal learning and retention. *Psychol. Rev.*, 1954, 61, 160-166.

- Postman, L. 95 (12), 126, 158 (31), 173  
 Pratt, C. C. 229, 232, 233
- Raiffa, H. 19 (6), 49, 262 (5), 269  
 Ribback, A. 239 (21), 269  
 Rosen, I. C. 145 (26), 173  
 Rosenzweig, S. 88 (13), 126  
 Ross, R. T. 238 (15), 253 (15), 269  
 Rossman, I. L. 137 (27), 173
- Saltzman, I. J. 142 (28), 173  
 Scates, D. E. 10 (3), 16  
 Scheible, H. 26 (19), 49, 158 (19), 173  
 Schneider, D. E. 161 (19), 172  
 Schoenfeld, W. 190 (10), 193, 258 (9),  
 264 (9), 269  
 Seeman, W. 88 (14), 126  
 Selfridge, J. A. 110 (10), 126  
 Sheffield, F. D. 145 (21), 172, 240 (16),  
 269  
 Shepard, R. N. 45 (20), 49  
 Shipley, W. C. 243 (12), 270  
 Singer, P. 30, 49  
 Skinner, B. F. 185, 187, 188, 194  
 Solomon, H. C. 24 (17), 49  
 Solomon, R. L. 145, 147, 158 (31), 173  
 Spence, K. W. 52, 83, 177, 181, 182, 188,  
 193, 194, 242 (24), 248, 252, 262  
 (25), 270  
 Spiker, C. C. 63, 84  
 Stafford, J. W. 177, 194  
 Stevens, S. S. 19 (22), 20 (22), 50, 81  
 (11), 84, 197 (17), 233
- Stone, C. H. 167 (23), 172  
 Suci, G. J. 45 (18), 49  
 Sumner, F. C. 155 (22), 172
- Taylor, J. A. 60 (12), 84  
 Thrall, R. M. 19 (6), 49, 262 (5), 269  
 Thurstone, L. L. 32 (23), 50, 207, 233  
 Tolman, E. C. 56 (13), 84, 183, 194  
 Tresselt, M. E. 157 (32), 173
- Underwood, B. J. 96, 101, 107, 126, 144,  
 154, 157, 164, 173, 185, 194, 259 (26),  
 270, 277, 292
- Van Zelst, R. H. 166 (37), 173  
 Verplanck, W. 190 (10), 193, 258 (9),  
 264 (9), 269
- Wallach, H. 262, 269  
 Wapner, S. 108 (17), 109 (17), 110 (18,  
 19), 126, 127  
 Waterhouse, I. K. 132 (10), 172  
 Waters, R. H. 53 (14), 84  
 Werner, H. 108 (17), 109 (17), 110  
 (18, 19), 126, 127  
 Whitfield, J. W. 259 (27), 270  
 Whittaker, E. 187, 194  
 Wickens, D. D. 243 (28), 270
- Young, R. K. 154 (38), 173
- Zander, A. 23 (24), 50  
 Zeller, A. F. 77, 84, 101, 127, 236, 270

- Gagné, R. 137 (17), 161 (16), 172  
 Galanter, E. 88 (14), 126  
 George F. H. 189, 193  
 Gibson, E. J. 137 (18), 172  
 Glanzer, M. 254, 269  
 Good, C. V. 10 (3), 16  
 Goss, A. E. 137 (27), 173  
 Graham, E. E. 57 (6), 83  
 Grant, D. A. 161 (19), 172  
 Greenblatt, M. 24 (17), 49  
 Greenspoon, J. 142 (28), 173  
 Grünbaum, A. 5, 16  
 Guthrie, E. R. 18, 49
- Haimowitz, M. L. 130 (20), 172  
 Haimowitz, N. R. 130 (20), 172  
 Hall, J. F. 29 (12), 49  
 Hall, M. 238 (15), 253 (15), 269  
 Harlow, H. F. 25, 49  
 Hebb, D. O. 229, 233, 249, 269  
 Helson, H. 246, 269  
 Hempel, C. G. 176, 193  
 Heyns, R. W. 23 (14), 49  
 Hildebrand, J. H. 3 (5), 16  
 Holton, G. 7, 8 (6), 16  
 Hovland, C. I. 145, 172, 238 (15), 253 (15), 259 (14), 269  
 Hull, C. L. 178 (13), 183, 188, 193, 223, 233, 238 (15), 253, 269
- Irion, A. L. 73 (1), 83
- Jahoda, M. 36 (15), 49  
 Jenkins, W. O. 240 (16), 269  
 Johns, E. H. 155 (22), 172
- Kanfer, F. H. 142 (28), 173  
 Kendler, H. H. 258, 269  
 Kerr, W. A. 166 (37), 173  
 Kessen, W. 230, 231, 233  
 Kimble, G. A. 230, 231, 233  
 Kneale, W. 178 (14), 193, 203, 233  
 Kobrick, J. L. 29 (12), 49  
 Koch, S. 189, 190 (10), 193, 258 (9), 264 (9), 269  
 Kohler, W. 262, 269  
 Korchin, S. J. 153 (2), 160 (2), 171  
 Krech, D. 228, 229, 233  
 Kreydt, P. H. 167 (23), 172  
 Kurtz, K. H. 138 (24), 173
- Levine, J. 24 (17), 49  
 Levinson, D. J. 54, 83  
 Levy, B. 157 (32), 173  
 Lewin, K. 132 (4), 171, 261 (19), 269  
 Lippitt, R. 23 (14), 49  
 Lumsdaine, A. A. 145 (21), 172
- McCandless, B. R. 63, 84  
 McCarthy, H. E. 3 (7), 16  
 MacCorquodale, K. 178 (16), 190 (10), 193, 223 (11), 225, 228, 233, 258 (9), 264 (9), 269  
 McGinnies, E. M. 105 (9), 126  
 MacKinnon, D. W. 178 (17), 189, 194, 225, 232, 233  
 Maier, N. R. F. 188, 189, 194  
 Marquis, D. P. 28, 49  
 Marx, M. H. 69, 84, 222, 232, 233  
 Maze, J. R. 258 (20), 269  
 Meadow, A. 24, 49  
 Meehl, P. E. 118, 126, 178 (16), 190 (10), 193, 223 (11), 225, 228, 233, 259 (9), 264 (9), 269  
 Melton, A. W. 281, 292  
 Mettler, F. A. 138 (25), 173  
 Meyer, D. R. 25, 49  
 Miller, G. A. 110 (10), 126  
 Miller, N. E. 248, 269  
 Minturn, A. L. 139 (8), 172  
 Montgomery, K. C. 63 (9), 84  
 Morant, R. B. 110 (18), 127  
 Mower, O. H. 24, 49  
 Mueller, C. 190 (10), 193, 258 (9), 264 (9), 269  
 Muller, H. J. 8 (8), 16
- Nagel, E. 176, 178 (8), 193
- O'Neil, W. M. 225, 233  
 Oppenheim, P. 176, 193  
 Oppenheim, S. 133 (7) 172  
 Oppenheimer, R. 13, 16  
 Oseas, L. 96 (11), 126  
 Osgood, C. E. 45 (18), 49
- Pap, A. 6, 16  
 Paterson, D. G. 167 (23), 172  
 Pennington, L. A. 53 (14), 84  
 Perkins, D. T. 238 (15), 253 (15), 269

# SUBJECT INDEX

Causality, and correlation, 33

Causation, finite, 6

Cause-effect, 35, 42

Concepts, causal identification of, 200; defined by experimenter behavior, 195; differences and similarities among, 202; elaboration of, 205; levels of, 195; operational definition of, 51, 195; phenomenon identification of, 198; postulated, 213; previous distinctions among, 225; reductive *versus* nonreductive, 226; response defined, 59, 204; summarizing, 223

Constant errors, 150

Construct validity, 119

Control group, definitional use of, 68, 75; failure to use, 130; inappropriate use of, 136; special problems of, 145; with systematic variation, 135

Counterbalancing, failure in, 108; nature of, 108

Definitions, literary, 54; operational, 51

Determinism, 4

Dimension, psychological, 21; quantification of, 41; unitary and complex, 43

Empirical explanation, illustrations of, 236; nature of, 235

Empirical integrator, 290

Empirical-postulational explanation, *ad-hoc* nature of, 250; illustrations of, 248; nature of, 247; verification of, 252

Environmental variables, confounding by, 128, 151, 159; generalizing from, 168; manipulation of, 36, 92, 128, 148

Equivalent groups, destruction of, 100; maintenance of, 92

Explanation, by models, 257; concept levels used in, 195; empirical, 235; ideal *versus* actual, 175; mixed empirical-postulational, 247; nature of, 234; physiological, 226; postulational, 253; problems in understanding, 177; purpose of, 180

Explanatory readiness, 175, 176, 186

*Ex-post-facto* experiments, 97

Factor analysis, 31, 80, 207

Freedom of inquiry, 9

Guessing, problems of, 158

Hypothetical construct, 225

Intervening variables, 225

Latin square, 108

Matched groups, forming of, 96, in *ex-post-facto* experiments, 97

Models, explanatory, 260; mathematical, 261; research, 258

Operational definitions, by response identification, 59, 204; by stimulus-response identification, 65; diagrammatic form of, 70; formulation of, 58; infinite regress of, 55; latitude allowed in, 54; levels of generality of, 55; priority in, 78, 239; provisional formulation of, 57; purpose of, 53

Operational identification, 205, 206, 235, 258

Operational validity, 222

Orthogonal design, 161, 279

# THE CENTURY PSYCHOLOGY SERIES

Richard M. Elliot, *Editor*

Kenneth MacCorquodale, *Assistant Editor*

- Social Psychology*, by Charles Bird  
*Learning More by Effective Study*, by Charles and Dorothy Bird  
*Psychological Counseling*, by Edward S. Bordin  
*A History of Experimental Psychology*, 2nd Ed., by Edwin G. Boring  
*Sensation and Perception in the History of Experimental Psychology*, by Edwin G. Boring  
*Readings in Modern Methods of Counseling*, edited by Arthur H. Brayfield  
*A Casebook of Counseling*, by Robert Callis, Paul C. Polmantier, and Edward C. Roeber  
*Beauty and Human Nature*, by Albert R. Chandler  
*Readings in the History of Psychology*, edited by Wayne Dennis  
*Techniques of Attitude Scale Construction*, by Allen L. Edwards  
*Modern Learning Theory*, by William K. Estes, Sigmund Koch, Kenneth MacCorquodale, Paul E. Meehl, Conrad G. Mueller, Jr., William N. Schoenfeld, and William S. Verplanck  
*Schedules of Reinforcement*, by C. B. Ferster and B. F. Skinner  
*Social Relations and Morale in Small Groups*, by Eric F. Gardner and George G. Thompson  
*Great Experiments in Psychology*, 3rd Ed., by Henry E. Garrett  
*Developmental Psychology*, 3rd Ed., by Florence L. Goodenough and Leona E. Tyler  
*Exceptional Children*, by Florence L. Goodenough  
*Physiological Psychology*, by Starke R. Hathaway  
*Seven Psychologies*, by Edna Heidebreder  
*Theories of Learning*, 2nd Ed., by Ernest R. Hilgard  
*Conditioning and Learning*, by Ernest R. Hilgard and Donald G. Marquis  
*Hypnosis and Suggestibility*, by Clark L. Hull  
*Principles of Behavior*, by Clark L. Hull  
*Development in Adolescence*, by Harold E. Jones  
*The Definition of Psychology*, by Fred S. Keller  
*Principles of Psychology*, by Fred S. Keller and William N. Schoenfeld  
*Psychological Studies of Human Development*, by Raymond G. Kuhlen and George G. Thompson  
*The Cultural Background of Personality*, by Ralph Linton  
*Vocational Counseling with the Physically Handicapped*, by Lloyd H. Lofquist



- Postulated concepts, growth of, 216;  
illustration of, 214, 219; levels of,  
213; purpose of, 213
- Postulational explanation, illustrations  
of, 253; nature of, 253
- Progressive errors, balancing of, 105
- Psychophysical methods, 41
- Randomization, of conditions, 108; of  
subjects, 93
- Random groups, errors in use of, 94;  
forming of, 93
- Reductionism, 126
- Reliability, need for, 22
- Research, analytical *versus* nonanalytical,  
271; idiographic *versus* nomothetic,  
88; pure *versus* applied, 8; reporting  
of, 289
- Research errors, nature of, 89; organization  
of, 91
- Response analysis, errors in, 160
- Response correlation, as research tool,  
29
- Response measures, inappropriate use  
of, 162; multiple, 27; nonequivalent,  
164; reliability of, 22; statistical analysis  
of, 160
- Scales of measurement, 18; nominal, 20;  
physical, 20, 41, 154; psychological,  
21, 41, 67, 154; rank-order, 21; ratio,  
20
- Scaling, 19, 41, 64
- Science, assumptions of, 3; purpose  
of, 1
- Scientific generalization, errors in, 166
- Standardization, 281
- Stimulus, active manipulation of, 36;  
natural variation of, 39
- Stimulus-variable elaboration, 205, 208
- Stimulus dimensions, quantification of,  
41
- Subject analysis, 46
- Subject biases, 149
- Subject variables, confounding by, 92,  
112; generalizing from, 167; manipulation  
of, 38, 46, 112, 159
- Systematic randomization, 108
- Task variables, confounding by, 148,  
154, 159; generalizing from, 169; manipulation  
of, 38, 92, 151, 154
- Technologist, 9
- Theory, and the scientist, 190; confusion  
about meaning, 177; purpose of,  
180 (*See*, Explanation)
- Two-stage experiments, confounding  
in, 152
- Validation, of tests, 29
- Variables, confounding of, 90; environmental,  
36, 92, 128, 148, 151, 159; instructional,  
38; intervening, 225; subject, 38, 46,  
112, 159; task, 38, 92, 148, 151,  
154, 159

*Studies in Motivation*, edited by David C. McClelland

*The Achievement Motive*, by David C. McClelland, John W. Atkinson, Russell A. Clark, and Edgar L. Lowell

*Current Studies in Psychology*, by F. Joseph McGuigan and Allen D. Calvin

*Principles of Applied Psychology*, by A. T. Poffenberger

*The Behavior of Organisms*, by B. F. Skinner

*Verbal Behavior*, by B. F. Skinner

*Diagnosing Personality and Conduct*, by Percival M. Symonds

*Dynamic Psychology*, by Percival M. Symonds

*The Dynamics of Human Adjustment*, by Percival M. Symonds

*The Ego and the Self*, by Percival M. Symonds

*The Psychology of Parent-Child Relationships*, by Percival M. Symonds

*Educational Psychology*, George G. Thompson, Eric F. Gardner, and Francis J. DiVesta. Also accompanying *Workbook* by the same authors.

*Selected Writings from a Connectionist's Psychology*, by Edward L. Thorndike

*Introduction to Methods in Experimental Psychology*, 3rd Ed., by Miles A. Tinker and Wallace A. Russell

*The Psychology of Human Differences*, 2nd Ed., by Leona E. Tyler

*The Work of the Counselor*, by Leona E. Tyler

*Experimental Psychology*, by Benton J. Underwood

*Psychological Research*, by Benton J. Underwood

*Elementary Statistics*, by Benton J. Underwood, Carl P. Duncan, Janet A. Taylor, and John W. Cotton. Also accompanying *Workbook* by the same authors.

*Persons and Personality*, by Sister Annette Walters and Sister Kevin O'Hara